1

IMPLICATION ANALYSIS: A PRAGMATIC PROPOSAL FOR LINKING THEORY AND DATA IN THE SOCIAL SCIENCES

*Stanley Lieberson**
Joel Horwich

Sociology and other social sciences struggle to emulate a model of scientific evidence that is often inappropriate. Not only do social researchers encounter special limits, but they are also handicapped by a distorted and idealized picture of practices in the "hard sciences." Ironically, while often obliged to use data of lower quality, sociology employs standards for evaluating a theory that are not attained in the hard sciences. After a brief review of these obstacles, we describe a set of procedures for using empirical data to rigorously evaluate theories and hypotheses without resorting

This research was supported by a grant from the Clarke Fund at Harvard University to Stanley Lieberson. Some initial parts of the paper were presented at the Seventh Annual Philosophy of Social Science Roundtable, Barnard College, March 11–13, 2005. We are indebted to organizers Alison Wylie, Paul Roth, and James Bohman for helpful comments. Also very helpful were the comments of two participants, Eric Schliesser and Julian Reiss. Christopher Jencks, Freda B. Lynn, Peter V. Marsden, S. M. Miller, and H. H. Winsborough made helpful suggestions. We are particularly indebted to Cathy Kenny and Sonya Keller for exceptional aid in the preparation of this manuscript as well as comments and advice on a variety of substantive and technical issues. Our development of Implication Analysis could not have occurred without a quotation in Rosenbaum (2002; 5–6) of a seminal paper by Cochran (1965). On then reading the entire paper, followed by the groundbreaking pieces using observational data to infer the role of smoking in lung cancer (Doll and Hill 1952; Cornfield et al. 1959), we began to develop the procedures described below. Direct correspondence to Stanley Lieberson at SL@WJH.Harvard.edu.

*Harvard University

to the mimicking of hard science. The interaction between theory and evidence normally involves deriving implications from the theory (usually referred to as hypotheses) and then ascertaining how closely the empirical evidence meets these implications. The appropriateness of the implications is a key factor in the entire operation, linking as they do the data and the theory. The evaluation of a theory is no better than the theory's implications (as generated by the investigator) coupled with the quality and appropriateness of the evidence. It is our impression, however, that because this step is insufficiently addressed, there are unnecessary problems in the evaluation of theories. We use the term "Implication Analysis" to describe our efforts to review and improve current procedures.

1. INTRODUCTION

Before turning to what we call "Implication Analysis," we begin with the central assumptions and observations leading to the development of alternative ways of using data to evaluate theories:

1. We assume that social processes are reasonably orderly and knowable and, in turn, that empirical evidence can help us evaluate theories that claim to explain these social processes.
2. Our data and methods are imperfect. Given the subject matters of sociology and other social sciences, the imperfections are often substantial—rather than merely modest departures from the ideal.
3. The linkage between theory and evidence is far more difficult than we want it to be (as is often the case with the way the social sciences are taught and the assumptions under which we operate), and the practitioner has to deal with this reality. Assuming we do not want to just give up on social science, we have to develop what may appear to be an oxymoron: a more rigorous approach using imperfect data and methods. Otherwise, we are *playing* at being a science of society—as opposed to *being* one.
4. On the one hand, it is almost certain that no single study will validate or invalidate a theory. On the other hand, there is a good chance that a large number of studies will not generate consistent results.
5. If the perfect is the enemy of the good, we want to figure out how to do research that provides as rigorous an approach as possible while recognizing the limitations that social research often encounters.

6. We embrace the role of researcher talent. Social science methodology is not about developing a foolproof system; it is at least partly an art. There is no mechanical system that will escape judgment, sense, and intelligence. But this does not eliminate the need for techniques and standards.
7. Our analysis of implications is stimulated by Fisher's brief comment on observational data, reported and elaborated decades later by Cochran (1965), and the elegant epidemiological research on the influence of cigarette smoking on lung cancer. Implication Analysis searches for as many consequences of a theory as possible and seeks in turn to find evidence to evaluate these consequences. There is a particularly appealing feature of this approach in which the full ramification of a theory is considered: It frees the researchers and theorists from worrying about whether everything "works." We rid ourselves of inappropriate notions of what results should look like if the theory is "true" or what it should look like if the theory is "false." And this should enable us to come to terms with our evidence and deal with the results in a more reasonable and appropriate way.

2. IMPLICATION ANALYSIS: A BRIEF OVERVIEW

Simply put: Evaluation in the social sciences commonly entails deriving implications from a theory (usually in the form of hypotheses) and then in turn ascertaining how closely the empirical evidence meets these implications.[1] We use the term "Implication Analysis" to describe our efforts to review and improve current procedures. Typically, implications are derived from the theory in order to develop empirical questions about the consequences of the theory. As such, implications are a central step in this process—a bridge between the theory and the data. The appropriateness of the implications is a key factor in the entire operation, linking as they do the data and the theory. However, the steps involved in going from theory to hypothesis are insufficiently addressed. Likewise, the widely understood difficulties in data analysis also create problems for generating an appropriate empirical evaluation. In other words, the ideal notion about how theory and data should interact is

[1] For further discussion of hypotheses, see Cohen (1934:202–7), Seawright and Collier (2004:289, 304), and Quine and Ullian (1978:66, 108).

just that—an *ideal*. Indeed, it by no means holds for all of the operations in the hard sciences as well.

Compared to tests of significance, there are relatively few rules and understandings about the generation of implications, their measurement, and in turn the evaluation of the theory. Are the implications true to the theory's intent? Are the data appropriate? What should be done when there are contradictory results? Or, put differently, how does it come about that contradictory results are obtained? Are there procedures that might help us decide? Are there some theories that are more appropriate than others for deriving implications? Obviously, this is a huge topic and entails various considerations, not all of which will be novel for any reader. Although we review various facets of this process, this paper should be seen as a starting point toward developing appropriate ways of analyzing implications and thereby improving our use of evidence. We will consider the logic of implications, the severe obstacles in appropriately evaluating these implications with less-than-ideal data, the problems due to the data we are customarily obliged to use, a more realistic and different metaphor for evaluating the evidence, and a variety of steps that should be considered in the evaluation of conflicting evidence. Our goal, in all respects, is to describe positive and useful procedures that should enhance the analysis of implications.

Ni Bhrolchain and Dyson (2007) provide a remarkable paper on causation in demography. Of special interest to us is a set of ten criteria for evaluating causal inference in the study of demographic change, as presented in Figure 1. It is included as having potential value as a set of implicit standards for evaluating not only a theory but also hypotheses generated from a theory. To the degree that the implication does not meet these standards, there is an increasing reason for questioning the implication and in turn the importance attached to the study. Of course, as we noted at the outset, there may be good reasons for the implication to not meet these standards. As already noted, there is a matter of judgment as to when such rules are inappropriate.

Figure 1 is an excellent set of standards to initially think about a theory—although a theory should not be rejected out of hand if there is failure in one of the standards. But we do want to comment about one type of theory—broad in its nature but very difficult to apply, let alone evaluate, in terms of its implications. Karl Popper is persuasive in questioning theories that include a series of chains or mechanisms leading to an outcome:

FIGURE 1.
Criteria supportive of causal inference regarding demographic change

1. Time order	The cause should precede the effect; or, where a process is cumulative, the start of the cause should precede the start of the effect.
	Simultaneous causation is possible in relation to mechanical processes.
2. Contiguity	The shorter the time between the cause and the effect, the stronger the basis for causal inference.
	Lags are possible, and may even be necessary, but they must be explained.
3. Duration	Causal inference is strengthened where the effect continues during the entire period in which the cause is operating.
	Not always applicable—causes of very short duration may have longer-lasting effects, and some processes may be irreversible.
4. Distinctiveness	Causal inference is more straightforward where both cause and effect are clearly differentiated and identifiable in a temporal context.
	True causes and effects may be hard to isolate from surrounding variability.
	Analogous to Bradford Hill's strength criterion, but distinctive effects may not be large and vice versa.
5. Direction	The effect should be in the expected direction—i.e., would the effect and its direction have been predicted before the event?
	Unexpected effects may occur and expected effects may be hard to specify.
	Analogous to Bradford Hill's plausibility criterion.
6. Proportionality	Causal linkage is better grounded when the scale of the effect can be considered proportional to the scale of the cause.
	Need not always apply—the criterion is subjective: apparently small causes can have major effects, and the reverse may also hold.

continued

FIGURE 1.
Continued

7. Recurrence	Causal inference is strengthened if the linkage occurs in a variety of settings. Context may, however, preclude exact replication. Not essential—some causes are historically unique. Analogous to Bradford Hill's consistency criterion.
8. No cause	Where the putative cause is absent, the effect is absent too. May not always apply in that multiple causes of a given event are possible.
9. Mechanism	To establish a causal link, a plausible set of intermediate links is required showing how the cause brings about the effect. Specifying and providing evidence of the mechanism involved is essential.
10. No alternative	All reasonable alternative explanations, including confounding, must be considered and ruled out. This criterion is simpler to satisfy where effects are large and distinctive. What is considered a reasonable alternative may change through time.

Source: Máire Ni Bhrolcháin and Tim Dyson, "On Causation in Demography: Issues and Illustrations." *Population and Development Review* 33 (1):25.

> . . . although we may assume that an actual succession of phenomena proceeds according to the laws of nature, it is important to realize that practically *no sequence of, say, three or more causally connected concrete events proceeds according to any single law of nature*. If the wind shakes a tree and Newton's apple falls. . .nobody will deny that these events can be described in terms of causal laws. But there is no single law, such as that of gravity, nor even a single definite set of laws, to described the actual or concrete succession of causally connected events; apart from gravity, we should have to consider the laws explaining wind pressure;. . .movement of the branch;. . .tension in the apple's stalk. . .The idea that any concrete sequence or succession of events (apart from such examples as the movement of a pendulum or solar system) can be described or explained by any one law, or by any one definite set of laws, is simply mistaken. There are neither laws of succession, nor laws of evolution. (1964:115)

If there is a mildly complex set of mechanisms, then the implications should be explicitly evaluated in appropriate piecemeal fashion. Otherwise, an implication will be almost certain to fail.

2.1. *Out of Context*

Virtually all theories in the social sciences can be shown to generate implications that will prove to be false. Why is this the case? First, most social theories are verbal statements, and no matter how precise the statements are, it is possible to take words and sentences out of context or out of the intended meaning and thereby generate an implication that is patently false. Beyond this, theories make assumptions that are not fully stated but are commonly understood. These assumptions are called *submerged statements*, "a set of conditions, limitations, definitions, contexts, and other modifications of what is literally being said that are—if communication is to work—understood by the receiver" (Lieberson 1998:186).

Consider how we might address an implication derived from a simple logical statement that does not require messy empirical evidence. If we start with a theory that all humans are mortal, then this would imply that all immortals are not human. How can this be? Clearly, the problem lies in the inherent difficulties that we have with language. To wit, the meaning of "immortal" is used to mean *not subject to biological death* rather than *someone with lasting fame*. This is obvious, and therefore someone deriving such an implication is applying a literal meaning without regard to the context of its intent. Probably most readers have encountered this type of situation, where someone makes a statement that is then taken out of context and thereby appears to be disproved (either because the context is not understood or just to be argumentative).

2.2. *Expansion*

This leads us to a sociological issue that is something of paradox. Let us start with two characteristics: the nature of the data set and the nature of the implication. If the conclusions are different between two studies that use the same data to evaluate the same implication, then most likely the paradoxical issue can be addressed through a tight examination of the research procedures. At the other extreme, where the two studies differ in both the implication derived from the theory and the data applied, it will take considerable effort to make sense of the outcome—particularly if either different hypotheses generate the same outcome or if the same hypothesis generates different outcomes. But, in any case, these are potentially manageable problems. It will be necessary to judge the appropriateness of the different hypotheses derived from the same theory. Likewise, it will be necessary to examine the differences in the data and procedures when there are different conclusions about the same hypothesis. In some cases, there will be clear problems about the implication or about the data or about the procedures, and this should be workable. In other cases, there will be no reconciliation between the research groups. Now what? Here we get into a sociological process. If the topic is an important one, ultimately others will be drawn into the problem and certain threads will begin to appear, such that the sources of the differences will be progressively understood. It is likely to stem

from outside parties. If the problem is important enough and lingers long enough, then a new cohort of social scientists will enter into the fray and the issue will be resolved.

Of course, it is heartening when the conclusions are the same—particularly when the data are different. Researchers may tilt in a certain direction about a given theory. And this may lead to using data for which the researcher has an a *priori* expectation that the chosen data will be especially likely to support or contradict the theory. Some of this danger, which can occur in good conscience, can be worked out by a process of reducing researcher bias as best as possible. Suppose the implications of a theory work very nicely for a given nation. The question is how does the theory work for several other nations? Implication Analysis requires that choice of additional nations be done in a manner to reduce a bias, to wit, with some form of randomization. In many cases, this is not possible because the quality of the data for different countries may be radically different. But a set of nations with suitable data should be generated and then randomly selected. The set from which the choices were made should be reported so that the reader can evaluate whether it suggests an inappropriate tilt. This notion of justifying the choice would apply to quantitative data sets as well—if more than one were reasonably appropriate.

We suggest that the researchers justify their selection of the data to evaluate the implication drawn from the theory, the outcomes, or, say, different countries, and so forth. But what if they are not? There are several considerations when different data generate different conclusions about a similar implication, or when the same data generate different conclusions about the same implication. The latter problem involves issues of replication. (What is the source of the different conclusions: statistical methods used? coding procedures? models? assumptions and so forth?)[2] If the parties responsible for these differences cannot resolve the problem, then there is a very nice social solution. If the issue is important, third parties will address it and—although it may also lead to a lack of agreement—over time there will hopefully be some consensus such that there is an understanding about how analyses of the same data

[2] We call the reader's attention to a discussion of this topic in the "Special Section on Replication and Data Access" of *Sociological Methods and Research* (2007, 36(2):151–219).

generated different conclusions and, in turn, which conclusion is more reasonable or—at least—if some truth can be found in both (the latter generating further understanding of the theory).

Ignoring the messy problems of data and evidence, several logical considerations are in order. If the evidence fails to support a theory, it does not follow that the theory is false. The implication drawn from a theory may have been incorrect, or the data of poor quality, or the bounds of the theory inadequately drawn. And, of course, if the results are based on only one or two cases, the evidence may be insufficient—if we recognize that the theory is a probabilistic theory. On the other hand, if the evidence supports the implications of the theory, we have to consider whether another theory would have generated the same outcome. Finally, a theory may be false, even if the results are congruent with the hypothesis (which is an appropriate implication). This is because the theory may be in error—even if in the specific case the outcome is correctly implied. These problems exist, but they should not be severe if the theory's implication is examined under a variety of contexts. In that circumstance, it becomes highly unlikely that a "true" theory will be persistently rejected or a "false" one persistently accepted.

Keep in mind, if the evidence for a theory is inconclusive, the absence of the affirmation does not necessarily mean that the theory is either false or valid—it means that we do not know. There are other plausible interpretations besides a poorly generated hypothesis that partially, but not fully, reflects the theory's implication. First of all, the data may be of poor quality—certainly a common issue. Also, the effect implied by the theory is drowned by the influence of contrary forces. Undoubtedly, there are many potential theories that cannot be appropriately evaluated for a variety of reasons: variables that are hard to measure in a suitable manner; social taboos; uncommon phenomena such that routine social research procedures will generate only a small number of cases.

Finally, it is important to distinguish between a *theory* and an *explanation* offered to account for a specific observation. Explanations may be based on a theory or simply suggest a theory, but no theory can be evaluated on its ability to account for an observation. Indeed it is difficult to provide suitable explanations for a specific event. If we ask why teenagers drop out of high school before graduating, there are a variety of valid explanations that can be offered. But if we ask why Mike or Jennie dropped out, the best that we can expect is a probabilistic

explanation unless the researcher knows an enormous amount about the child. And, even then, it would be a probabilistic statement. Indeed, this is probably a major difficulty that sociology and social science—more generally—face: an expectation from the public that is actually, in its specific nature, very hard to achieve.

2.3. *Limits*

It is one matter to believe that there are theories that will operate throughout time and space. We do not have to worry about the first belief since we can accept the theory as true, but essentially irrelevant in a vast number of conditions where its influence is virtually nil. But it is another matter to believe that there are all-powerful theories operating with such force that they will make their presence felt regardless of countervailing conditions. The latter is unlikely. As a consequence, we become interested in the limits of a theory. Under what circumstances can we expect it to operate such that we can observe its implications and, in turn, under what circumstances can we expect its influence to be muted such that implications are for all purposes absent?

In either case, we need to know the conditions under which the implications of a theory might be expected to operate, in the sense of generating observable outcomes. And, of course, in knowing these limits, we also know where the implications do not appear—or if we visualize this as on a continuum, where the impact becomes progressively weaker.

Most theories do not state the necessary conditions under which the implications of the theory are relevant or visible. Put another way, we start with the belief that a theory does not always "work," in the sense that its implications do not always occur, nor do we have reason to think they should always occur. This reflects not just the usual issue for a probabilistic theory (a matter that we certainly accept as one reason for inconsistency in observations), but the implications will fail to appear in the absence of necessary conditions. It is easy to see this in any theory with an important quantitative dimension. For example, does alcohol consumption among teenagers affect their attitude toward school? We may have a theory that it does but not find that the theory matters for students who rarely drink. The theory has implications that are generally true, but there are limits to its effect. This is easy to see because

it is a quantitative theory. The point, though, is that we can expect theories to have conditions under which the implications operate and other conditions under which they will not operate. Learning the limits through empirical study is all right, albeit sluggish, and also may lead to artificial *ad hoc* bounds that have no rhyme or reason. In short, theorists should make some effort to estimate the theory's limits. But if they fail to do so, it is necessary to keep in mind this almost certain limitation.

3. BACKGROUND: THE REAL WORLD OF SOCIAL SCIENCE

Admittedly, social science data often have special problems. But the standards and procedures used to link the data with theory are often *unnecessarily* shaky and unconvincing—even after taking into account these difficulties. This leads to the somewhat counterintuitive notion that current standards of evidence will reject theories that should be accepted and those same standards will lead us to accept theories that should be rejected. Note: the focus here is with *epistemological* errors caused when conventional procedures are inappropriate for connecting data with the theory. This is different from Type I and Type II Errors generated in significance tests; we are not concerned with this important and well-understood *statistical* issue. Likewise, in discussing theories that *should have been rejected* or *should have been accepted*, we concur with Hill (1965:300): "all scientific work is incomplete—whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge." Again, we are not discussing an erroneous conclusion due to observations and theories that were unknown at the time. These errors should be self-correcting when new information is available. Rather we refer here to erroneous conclusions unnecessarily caused by the use of improper standards in the analysis of existing data.

Why do we expect contemporary standards to often reject valuable theories and accept useless theories? Our point is illustrated by two radically different examples. First, visualize a widely studied theory, with extensive research data. There is no problem if the results either consistently support or consistently reject the theory. However, for reasons that we develop later, there is a good chance that not all of the results will generate a consistent conclusion. In the absence of consistency, a

variety of conclusions is possible. For theorists and researchers who operate with the mistaken expectation that an empirical "test" should provide a clear answer about a theory (hopefully a small number), then the conclusion would be that the theory is falsified because the theory cannot explain some of the results. If so, should a new theory be proposed that will incorporate all of the results? Or, in order to support a favored theory, should contradictory evidence be ignored? Or should the solution be based on the majority of the results, such that the theory receiving the most empirical support is judged to be "correct"? Without an understanding of the standards to be followed, a case can be made for any of these conclusions. We observe contradictory results and do not know which conclusion may appear to be more justifiable than others. The evidence does not provide as much information as we have a right to expect, granted that the data are almost certain to have some problems. The set of procedures that we follow is incomplete. The ambiguities raised by contradictory results are often swept under the rug—thanks to the "theory testing" broom such that any contradictory result is taken to mean that the theory is wrong. In any case, we simply have no clearly stated procedure. Although a rigid rule might itself be damaging, somehow we should have more of an *a priori* understanding of what is required.

The second example stems from the opposite situation: The empirical evidence is strikingly consistent in either supporting or contradicting the theory, but it is insufficient to justify a strong conclusion and, at best, the results are promising. Here the problem occurs when the results are *over interpreted*. To understand the importance of this caution, consider the constant flow of pharmaceutical studies that report an initial positive outcome for a new drug but then are reversed after further testing and analysis. In short, these two radically different situations illustrate our assertion that sociology—and the social sciences in general—have failed to develop rigorous standards for evaluating the results.

Moreover, a void in epistemological rules creates an ambiguous situation that can subtly tip the interpretation of the evidence toward the investigator's beliefs and dispositions. The outcome often resembles a debate, with "telling" points and observations tossed back and forth and the conclusions becoming a function not only of the strength of the evidence but also of rhetorical skills. (For a discussion of the disconnect between theory and evidence in economics, see McCloskey 1998). This is

discouraging for both social scientists and consumers of social science—particularly for those who share our assumption that social processes are orderly and knowable.

4. SPECIAL OBSTACLES IN THE USE OF EVIDENCE

We briefly review three particularly important obstacles in the use of evidence in evaluating theories:

4.1. *A Counterproductive Model of the Hard Sciences*

Sociology and other social sciences struggle with a distorted image of the hard sciences that is not only an inappropriate model but includes standards that are even more stringent than those used in most of the hard sciences. The image of hard sciences employing a straightforward *test* that leads to a hard and firm evaluation of a theory's validity is false and harmful for social science. Yet the language of sociology is full of this hard science imagery. Sociology, for example, employs such notions as "testing" a theory to see if it is true; or "controlling" for various conditions in evaluating a theory, as we might do in dealing with manageable inanimate objects or biological specimens. There is the expectation that a theory—if it is valid—is able to "predict" future and unknown events. Indeed, prediction is a test of the "power" of a theory. Even more complex is the expectation that we should be able to *choose* between competing theories simply by finding a situation where they predict different outcomes. (It is never quite clear, by the way, what to do if neither predicts correctly, let alone if Theory A does better in one test and B in another.) When the two theories correctly predict the same outcomes repeatedly, we in turn employ parsimony to choose between them, again following the solution we attribute to hard science.

Ironically, the hard sciences rarely use such a simplistic standard; a conclusion about the utility of a theory rarely stems from a single study or even several studies. Perhaps this image originates from the way natural science is (or was) taught in elementary and secondary schools—as if one experiment could lead to the rejection of phlogiston theory; or another experiment could toss Lamarck's theory of inheritance of

acquired characteristics into a garbage can. In practice, it can be difficult to simply generate the necessary data to evaluate a theory, let alone have sufficient confidence in the results. Witness, for example, the obstacles faced in properly comparing the prediction derived from Newtonian physics with the prediction stemming from Einstein's general theory of relativity (Lieberson 1992). To be sure, the set of worldwide astronomical observations based on the eclipse of 1919 favored Einstein's prediction over Newton's (Mayo 1996). However, the conclusion involved more than simple comparisons between the observations and the expectations generated by the theories—indeed these results alone were disputed and reinterpreted by some who questioned whether Einstein's theory was the cause of the reported observations. According to Mayo's review of the evidence, the matter was settled because Einstein's theory could explain not only the astronomical observations but also Kepler's third law, *and* the motion of Mercury's path.

In similar fashion, the interplay between theory, implications, and evidence is more casual in the hard sciences. In practice, prediction in the hard sciences is often different from, and less rigid than, the social scientist's tighter view of prediction (Brush 2003). The role of intuition is more openly acknowledged in the hard sciences and the distinction between deduction and induction is much more fuzzy. Formal rules of evidence such as those described by Mill (1874) are seen as relevant only for the early stages of a science (see Holton 2004) In any case, the assumption of a tight linkage between theory and evidence makes no sense for the social scientist who typically must deal with an enormous range of conditions that not only may have an impact on the outcome but may even overpower the consequence anticipated by a specific theory. (And the data are sometimes problematic, to boot.) Consequently, many of the procedures social scientists use for empirical evaluation of theories are inappropriate.

4.2. *Excessive Reliance on the Experimental Model*

Given the view of experimental evidence as the "gold standard" for evaluating and testing theories in the natural sciences, it is not surprising that the experimental model profoundly affects expectations about the way data and evidence should operate in the social sciences. The language of experimentation is often employed even when true experimentation

is not possible for ethical or practical reasons. There is a widespread tendency to employ various statistical procedures to approximate the conclusions that an experiment would have produced. This is understandable and is often successful. But it is problematic on two important counts. First, many natural sciences do very nicely without relying *exclusively* on experiments (for example, evolution, meteorology, many aspects of astronomy, and geology). Observational data play a major role in generating a high level of confidence in the evaluation of a theory. Witness, for instance, the important role observational evidence played in the comparison between Einstein's and Newton's theories (albeit recall from our earlier discussion that these astronomical observations based on a solar eclipse were not the only basis for favoring Einstein). Or consider Darwin's reliance on observational data in developing and expounding the theory of evolution (Lieberson and Lynn 2002). The potential for such rigorous knowledge is both underappreciated and underutilized in social science. However, the use of observational data should not be confused with the questionable disposition to appropriate Mill's methods to social processes in a helter-skelter fashion.[3]

Since true experiments are simply undoable in many areas of social research, alternatives are often necessary. But even when randomized experiments appear to be possible, the results are often compromised. For instance, randomized experiments employed to evaluate drugs, procedures, and diets often generate inconsistent and misleading results. These obstacles tell us a lot about randomized experiments in the social sciences because the randomized health studies usually benefit from better financing and are probably easier to perform than most social experiments. Among the obstacles observed by Nowak (1994) in her review of clinical trials are the following: The number of subjects are too small to detect treatment differences; subjects drop out before the study is completed, thereby complicating the possible inferences; there is an absence of true randomization because experimenters fail to follow the indicated protocol (to say nothing of using an improper protocol); there is improper inclusion of subjects for whom the experiment is irrelevant or improper, which impacts on the estimated magnitude of the test

[3] Indeed, in a section widely ignored by practitioners of these methods, Mill (1874:608–13) demonstrates the inappropriateness of the Method of Difference, the Method of Agreement, and the Method of Residues for understanding social processes.

effect; populations of subjects are non-randomly chosen either because of exclusion rules *or* because of non-random assignment to the test and control population; and pressure is exerted by subjects in the control population who demand to be included in the test group because of perceived advantages. (By the way, this is a particularly striking matter in social experiments—for example, when parents discover that other children are in smaller classes, with more advantages in equipment, extra programs, additional teachers, and the like. These parents then press administrators to have their offspring moved to the advantaged setting or—if need be—sent to a different school. In any case, the randomization model underlying the experiment is damaged. Indeed, such a realization undercuts the control/comparison since the performance of the children is no longer a comparative counterfactual test.)

There are additional problems in medical research that are relevant for evaluating social experiments—even if randomized. The analysis of an experiment's results into subgroups is tricky; on the one hand, there is no *a priori* reason to assume that a given test condition will operate in a uniform manner for all subsets of the population. It is reasonable on that score to consider these differences. On the other hand, Nowak (1994) noted a disposition to break down the test subjects into progressively smaller subsets that eventually generate differences that are as much likely to reflect random outcomes across small populations as to reflect a meaningful result. Aside from the over-analysis of subgroups, Nowack also found questionable surrogate markers for clinical endpoints; indiscriminate inclusion in a five-year survival study of patients who had taken less than four-fifths of their medication and hence should have been dropped from the data used to establish the drug's effect (this becomes sticky because there is also an argument against dropping them as well as an argument against keeping them); an inability in some medical research to sort out side effects of a drug—in one case, for example, the drug was beneficial by reducing patient's high blood pressure but alas was also toxic, thereby nullifying its positive effect. This is interesting to us because in the social sciences it is relatively rare to see a consideration of "side effects"—to wit, research that considers not merely the targeted effect of a given program but also other effects of the same program. Appropriately, we would condemn medical research that did not examine side effects, but rarely do social researchers consider side effects for social policies. Another outcome in randomized medical research was a result that proved to be very

effective for treating colon cancer among young females in one trial but was specially effective for older males in a second trial one year later. This radically different conclusion about a drug's effectiveness appeared to reflect the small size of the trials coupled with a short-lived benefit. Again, we see the gold standard randomized experiment is not quite the cure-all for social research (see also Harrington [2000]).

It is all the more dangerous to assume that the statistical manipulation of observational data yields a reasonable estimation of what a true experiment might have yielded. As we have seen, a randomized experiment is *itself* a counterfactual statement about what would have happened to a population if a test condition had not been administered to it (say the administration of a drug, the development of a program for preschool education of children from impoverished homes, training programs, all sorts of government policies, and the like). The conversion of non-experimental data into an estimation of the outcome generated by a true experiment is, at best, a counterfactual estimate of what the results would have been under a true experiment. In point of fact, the observed data are likely not to be the same data as would have been obtained had the study been designed as an experiment. While naturally occurring experiments—events that have the crucial form of an experiment but were not induced by researchers—overcome some of these obstacles, relatively few of the questions of interest to social scientists are likely to align with such rare events. (Bear in mind that experiments can be plausible and highly desirable instruments for examining theories and hypotheses; also, counterfactual statistical estimating procedures have their potential place as well. Our goal is not to deny the promise of these procedures, but rather to note that we should not automatically defer to the evidence they generate. Actual experiments—let alone counterfactual experiments—are not going to provide a full solution to the data issue.) Granted, a well-executed randomized experiment provides the social researcher with a strong basis for causal inference; but even then, a second issue is the broad range of possible conditions that operate to affect the specific results from such an experiment. For example, in the case of the effect of a training program, the experiment can tell the researcher what the effect is of the *specific* training program on the *specific* subjects in a *specific* location. A wide variety of experiments would be needed to work out these conditions. And this becomes increasingly difficult when dealing with a counterfactual statistically adjusted design.

4.3. *Less Than Ideal Data*

In passing, Sections 4.1 and 4.2 have touched on some of the problems due to the nature and quality of social science data. It is often difficult for social researchers to obtain high quality measurements of the dependent variable, let alone causal factors. (Consider, for example, such matters as subjects dropping out of a study, answers that are incomplete, comparability over time and place, events that are not well-documented, and distortions by respondents.) These "obvious" considerations are apt to be skipped as the researcher prematurely focuses on the theoretical relevance of the results. However, it is relevant here because efforts to create an approximation of the results that would occur were the data of better quality should not ignore the need to create more appropriate standards that are suitable for evaluating the questions with the available data. Ignoring the latter is the social science equivalent of telling a story about an operation that is a success but the patient dies. In this case, the form of the analysis is wonderful but the data are not suitable for the problem.

4.4. *In Short*

Little is gained when theories are evaluated by combining unrealistic standards with less than ideal data, and employing methods that rest on unrealistic assumptions. Worse yet, this ritualism prevents us from recognizing the possibility of a more appropriate way of using data to evaluate theories. Our proposals are not a strict recipe or formula to be followed; there is no mechanical system that will escape judgment, sense, and intelligence. Indeed, one of the mistakes commonly made in contemporary social research is assuming the existence of standards or procedures that are applicable in a mechanical way to evaluate data (as, say, variance explained, predictive power, tests of significance, and the like). Social science methodology should not be viewed as developing a foolproof system in which data are plugged in at one end and the best answer is generated at the other. As is the case in all areas where evidence is gathered to evaluate a theory, success is often marked by imaginative and creative efforts. This is at least partly what might be called an "art." The general idea entails a search for relevant data (based on a far wider view of what is typically seen as *relevant*) coupled with appropriate

techniques and standards that can be used to squeeze every possible bit of use for the problem at hand.

5. IMPLICATION ANALYSIS: FISHER ON OBSERVATIONAL DATA

The inspiration for our proposal stems back to a conversation in the mid-1930s between the great British statistician R. F. Fisher and another distinguished statistician, W. G. Cochran. Two decades later, Cochran (1965) recalled the conversation:

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many *different* consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. (P. 252)

Cochran illustrated Fisher's advice with examples from the epidemiological triumph of that period—the extraordinary success in using observational data almost exclusively to determine the influence of cigarette smoking on health. Two types of studies are both prominent in his comments and of special interest to us. The first, alluded to above, expands on the theory under consideration. In the case of smoking, this means doing more than simply comparing the death rates of smokers and nonsmokers. It means comparing a wide variety of conditions: persons who smoked different amounts for comparable lengths of time; smokers who smoked the same amount but started at different ages; ex-smokers who had stopped recently and those who had stopped earlier; and so forth. Cochran (1965:252) concludes

> Of course, the number and variety of consequences depends on the nature of the causal hypothesis, but imaginative thinking will sometimes reveal consequences that were not at first realized, and this multi-phasic attack is one of the most potent weapons in observational studies. In particular, the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others.

The other type of special interest is the generation of possible alternative conclusions that might explain the observed events *without reliance* on the theory under study (in this case, the smoking hypothesis). Note that this is a different step than the customary use of tests of significance to examine a broad null hypothesis—here it is to examine a specific alternative causal theory.

Cochran's comments and his examples from research on smoking are of more than historical interest: They are relevant to present practices. He is responding to situations where the data are largely observational and—even then—often of mixed quality, and where true experimentation is rare and the studies are often statistical estimations of true experiments. Cochran (1965) describes the challenge of drawing conclusions based on such data:

> The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis. If some results appear to support the hypothesis, some contradict it and some are neutral, reaching a verdict demands much skill. Obviously, the investigator should consider whether some revision of his hypothesis will remove the contradictions. In default of this, he cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance. (Pp. 253–54)

Regrettably, Cochran's observations about data are still highly relevant today. The present-day data situation in much of the social sciences is not much more superior to that confronted by Cochran decades ago. To be sure, the data sets and surveys are of better quality (and more numerous), and computers certainly allow for the manipulation of data and the use of elegant new statistical procedures, but researchers are still obliged to work with one form or another of observational data. Moreover, it is by no means certain that contemporary methods provide us with pathways to a high level of confidence for using the data to evaluate theories.

This brings us to the starting point of what we call *Implication Analysis*, which describes three basic steps in evaluating a causal hypothesis. First, the implications of a theory are developed for all sorts of conditions and circumstances (for example, different groups, different times and places, different social conditions) and these implications are then empirically examined. In all cases, studies will ask if the expectations based on the theory seem to "work."

The second step considers whether the evidence supports an alternative theory, again generating implications in a variety of contexts. Here we ask if there is evidence suggesting that the initial causal hypothesis is not causing the observed outcomes. In other words, is there an alternative that is even more successful in accounting for the observations covered by the initial theory and can extend to even broader ramifications? Or, at the very least, is there an alternate theory that covers important events that the initial theory cannot handle? There certainly will be instances when the conclusion is very clear after following these two steps. But moving away from an ideal image of how science should work and what data should be like, in reality more than likely we will confront somewhat mixed results with respect to the initial theory and also the alternative accounts.

If neither the initial hypothesis nor the alternative is fully satisfactory, a third step is needed to evaluate and compare the evidence supporting the initial theory and the alternative causal accounts. This step differs radically from the current disposition to either search for a third theory or at least somehow pool the two theories when confronting inconsistent results in which theories seem to work somewhat but not completely. Here our earlier observations about the quality of data evidence come into play. Rejection is a premature step unless one has gained sufficient confidence that it is not the evidence that is deficient

but rather the theory. This is a treacherous and challenging task: There is the danger that the theory will be erroneously rejected or accepted because of the poor quality of the empirical evidence. We will propose a set of checklists for evaluating such contradictory evidence.

6. EXTENDING A THEORY

In the social sciences, a useful theory should suggest a wide array of expansions. Some theories are narrow and read as if they are nothing more than a slightly modified restatement of the research finding. Other theories are broad and far-reaching. In either case, they should be approached in terms of gauging their range and potential value. (An extremely narrow and non-expansive hypothesis is likely to be replaced by a hypothesis that will incorporate these observations as well as have other applications.) Theories stated in broad and far-reaching terms need to be considered in all of their broad and far-reaching consequences in order to be appropriately evaluated. In any case, the initial examples presented of where either type of theory "works" or "fails" are nothing more than *examples*—even if they are attractive. As Rogowski (2004) notes, "inference proceeds most efficiently by three complementary routes: (1) making clear the essential model, or process, that one hypothesizes to be at work; (2) teasing out the deductive implications of that model, focusing particularly on the implications that seem *a priori* least plausible; and (3) rigorously testing those least plausible implications against empirical reality" (p. 76).

The typical research procedure starts with a set of observations that we would like to understand. We then apply a theory or hypothesis to see how well it handles these observations. Or the direction is reversed such that evidence is gathered in order to evaluate a theory. In either case, controls are also applied to increase our confidence that the apparent utility of the hypothesis is not due to other factors. If the hypothesis does seem to operate fairly well, we are done with the task. Another conventional reason for linking data occurs when we seek to determine if a theory is "true" or "false." We think of some data set that might be used as a "test" of the theory and then see if the theory meets the test (again applying a set of controls). As indicated earlier, it is inappropriate to assume that any one empirical study is sufficient to reach a true-false conclusion. Instead, following Fisher and Cochran, we propose

and illustrate more appropriate and demanding applications that do not rely on any single event, study, or data set, but rather explore the hypotheses generated by the theory and then, in turn, bombard these hypotheses with evidence. The goal is straightforward: We should search for as many extensions and consequences of the theory and then evaluate these consequences by obtaining or developing relevant data. We are not usually oriented to thinking this way.

A more demanding test occurs when a theory is examined under circumstances that differ from the context in which the theory was initially developed. Here, the theory is less likely to "pass," but "failure" does not necessarily mean that the theory is wrong for the original context or for contexts that are similar to it. In most social science situations, there are so many conditions operating to potentially affect the outcome that it almost takes a mutually shared suspension of disbelief to accept a conclusion without demanding not only replication but also wide-ranging studies that expand beyond the initial results. We speculate that one reason for the reluctance to examine the bounds of a theory is the mistaken notion that a "negative" result along the way will lead to the rejection of the theory—its *falsification,* as it were—if not by the investigator, then certainly by critics. Because a low threshold for rejecting a theory is so inappropriate in many areas of social science, we develop below some important steps for evaluating inconsistent evidence without simply tossing out the theory. The paradox of such an expansion is that it will almost certainly lead to "negative evidence" where the implications of the theory fail to hold, but a useful theory will in fact have ramifications that move beyond the very literal question under review.

6.1. *Bourdieu Meets Data*

Sometimes the extension of a theory is simple yet very effective. By examining Bourdieu's (1984) theory of the role of cultural capital in contexts differing from the one in which it was developed, the limitations and boundaries of this theory are explored. Erickson (1996), for example, used a work setting to examine the ramifications of Bourdieu's theory in establishing and maintaining hierarchies. Her results are strikingly different from what would be deduced from a simplistic reading of Bourdieu. To be sure, cultural capital played a conversational role in her analysis, but it was hardly the type of capital discussed by Bourdieu.

Likewise, Halle's study (1993) of class factors on the display of art in New York homes, as well as class differences in attitudes toward different types of paintings, is relevant to our discussion here. On the one hand, the class differences in attitude and display were in keeping with deductions we might make from Bourdieu, but a striking deviation is found. Halle found that the art displayed by the higher SES subjects does not reflect their cultural capital or their ability to understand the work but rather their notions of what would be appropriate and attractive art in their home. Finally, Lamont (1992) examined the role of cultural capital among respondents of a comparable class in the United States and France, finding radical differences between the two. The lesson is clear: Bourdieu's theory is a "big" theory, with wide-ranging applications and consequences. It is unlikely to always fail, but it is also unlikely to always hold. Therefore, the examination of the value and utility of the theory can be helped by drawing consequences from the theory and applying them to a wide variety of contexts. It is not helpful for a study to conclude that his thesis is "false" or "wrong" because it fails to hold in a specific empirical context. Nor are we helped by a study claiming to prove that Bourdieu's theory is correct in some absolute sense because the results agree with his theory. In effect, bounds need to be drawn around the theory that reflect the preponderance of evidence on different facets of it. On this score, we admire Erickson for avoiding the temptation to conclude that Bourdieu's theory is wrong—obviously, her study should generate a wide number of different studies of cultural capital used in the workplace before such a conclusion would be appropriate. More likely, it will lead to an analysis of the work contexts in which high cultural capital may be enhancing, neutral, or actually counterproductive for those who display it. Put another way, exploring the role cultural capital plays in various contemporary Western settings—to say nothing of non-Western settings—or earlier periods would almost certainly further help to mold and expand Bourdieu's theory. Obviously, no single scholarly group can do all of this; and that is why it is appropriate to think of human knowledge as a communal effort, particularly when there is either a theory or empirical result that has the possibility of deeply impacting a potentially important problem (we will return to this later). Many extensions are not implied by the initial theory. For example, there is nothing in Bourdieu that would lead us to view him as describing the United States. But it is still useful to search for the boundaries of the theory.

6.2. *A Theory About China Examined with Russian Data*

Nee and his associates theorized that the transition from state socialism to a market economy would increase returns to human capital and entrepreneurial initiative and decrease returns to political position (Nee 1989, 1991, 1996; Nee and Matthews 1996). Gerber and Hout (1998) sought to expand the applicability of market transition theory by applying it to another important case, Russia. Their procedure illustrates the potential of the general methodological model we have in mind. Based on their reading of Nee's work, Gerber and Hout generated six general hypotheses and 16 specific empirical hypotheses to be tested with five years of survey data. For example, from the general hypothesis "Market transition increases returns to education," three specific expectations were developed:

1. The total effect of education on earnings in Russia increases from 1991 to 1995.
2. At any given time, the returns to education in the private sector of the economy exceed the returns to education in the state sector.
3. The returns to education in the private sector in 1995 exceed the returns to education in the private sector in 1991 (Gerber and Hout 1998:7–8).

After examining the 16 hypotheses, Gerber and Hout concluded that the predictions of market transition theory were contradicted by the Russian data. They argued that the case of China, which Nee used to develop his market transition theory, could not be applied generically to other countries. The key differences between China and Russia, they stated, were the pace at which market transition had taken place as well as the political origins of each country's transition. Russia began market reform after the collapse of the Soviet political system, which resulted in "steep contraction, rather than economic growth," whereas China's transition had been gradual and simultaneous with "enforced political stability" by the Chinese government (Gerber and Hout 1998:36). Therefore, they suggested, the social stratification implications of the move from socialism to capitalism were more dependent on country context than had been suggested by Nee's theory. Nee later argued that a certain number of Gerber and Hout's hypotheses, particularly those dealing with predictions of decline in the gender gap and differences in

benefits to actors in the private sector versus state sector, were "not truly *derived* from market transition theory" (Cao and Nee 2000:1186). He in fact argued that Gerber and Hout had treated "economic depression, high levels of unemployment, and hyperinflation as processes intrinsic to market transition," but, in his estimation, the presence of these processes might indicate that Russia had not established a market economy at all during the years of the Gerber and Hout study (Cao and Nee 2000:1186–87). However, the ensuing development of the field suggests that Gerber and Hout's work was part of a body of research that led to a more nuanced perspective on market transitions (see Szelenyi and Kostello [1996]; Nee and Cao [1999]; and Cao and Nee [2000]). Without taking a substantive position, we would argue that Gerber and Hout's approach—generating numerous implications from a theory and examining them systematically—contributed to a debate over market transition theory. As Zhou (2000:1190) argued, empirical data suggesting a possible contradiction with market transition theory provide an opportunity for theoretical growth. This approach to evaluating theories through close examination of their implications encourages ongoing refinement of understanding of sociological phenomena. In this particular case, the arguments back and forth, with the addition of new participants (i.e., third parties), are an ideal way of evaluating and specifying the suitability of the implications developed and the quality of the data.

6.3. *Expanding the Settings*

The great wave of European immigration to the United States, culminating in the immigration restrictions enacted in the 1920s, was marked by an important difference in the arrival timing of different groups. Up through about 1880, European immigrants were primarily from Northwestern Europe (the "Old Groups"); after 1880, migrants from South, Central, and Eastern Europe were the major sources (the "New Groups"). There was considerable public and scholarly debate on the relative success of these two groups, with the latter's poorer standing often attributed to biological factors and character flaws. Forgetting the questionable explanations, was this even a valid analysis of their different levels of accomplishments? Length of stay in the United States, along with the existence of earlier members of the group in the country, made it difficult to compare those members of the Old Groups coming

after 1880 with New Groups who were migrating at the same period.[4] Although there are numerous ways of dealing with this, there is a problem in the sense that the timing of the waves are different and therefore it is hard to sort the generational effect from the historical set of conditions—a problem that continues to occur in the United States for relatively new immigrant flows.

Australian immigration policy after World War II provided what we would now call a "natural experiment" that can be used to address the key features of the comparative issues in the United States. Prior to the end of World War II, Australia was open to immigration from the British Isles but largely closed to immigration from anywhere else in Europe. After World War II, they opened up to immigration from everywhere in Europe—indeed, they encouraged it. Consequently, immigrants from Italy, Greece, Yugoslavia, and 16 other "New Groups" were first reaching Australia in sizable numbers at the same time that northwestern Europeans from nine nations (including Denmark, France, Germany, and the Netherlands) were arriving. According to Australian data (Lieberson 1963), these groups were compared over a wide range of measures: six work force characteristics; three spatial distribution measures, including one that covered residential segregation in four metropolitan areas; intermarriage, naturalization, and birth rates; and three measures of social problems. The conclusion was straightforward: There was no evidence to support the broad theory of Old Group superiority developed for immigrants around 1900 in the United States. Is the United States the same as Australia? Are the groups coming after World War II identical to the same groups arriving in the United States in sizable numbers at least 50 years, and in some cases centuries, before? Obviously no. However, at the very least, the results are harmonious with an alternative interpretation of events in the United States. In this case, there are grounds for reconsidering the initial theory about Old-New differences, but the conclusion drawn from Australia is by no means any more definitive than the conclusion drawn in the United States.

All three of these examples, although taking different forms, are based on a common assumption—namely, a theory cannot be

[4] This is a brief summary of a complex and important issue at the time. E. A. Ross, a major sociologist in his time, and a president of what was then called the American Sociological Society, wrote a book in 1914 expressing concern with the racial and cultural deterioration of the United States; for some choice quotes, see Lieberson (1980:25–26).

adequately evaluated using a single data set. Whether the initial data set supports or fails to support the causal theory, it is almost certain to be insufficient evidence. Rarely will there be a single study that is hardy and extensive enough to nail down the utility of a theory (except perhaps for what we might call *pseudo-theories*—namely, when a finding is restated in the form of a narrow and highly restricted causal proposition). Moreover, there is no reason to assume that the impact of a theory—even if framed as a "universal theory"—will be obvious in all times and places. This is because the influence of other conditions may modify and outweigh the impact of the so-called universal theory. In addition, there is no reason to assume that it is possible to successfully "control" all of these other influences.

By considering the implications of a theory in broader terms and asking if the relevant evidence supports this theory, we are evaluating the theory as well as considering the bounds under which the theory holds. This is essentially the procedure endorsed by Schliesser (2005) in his review of the underpinnings of experimental economist George Smith (1994, 2002). Theory is not viewed simply a source for predictions—rather it is viewed as an engine for discovery (Schliesser 2005:51). The goal is not confirmation or refutation—as we might think of it in the falsification issue—but rather it is the generation of information to improve and modify the theory. If we do not expect that all possible ramifications of the theory must hold, the interaction between evidence and the causal implications of a theory will be less timid and bolder in considering more realistic patterns.

As we shall see, a negative result does not mean that the theory is inherently flawed. On the other hand, a result that is consistent with the theory is *promising*, but its utility is hardly established. This elaboration means generating as many consequences of the theory, and from as many sources, as possible and evaluating the body of evidence. It also means that it may sometimes be necessary to use less-than-ideal data (of course, that is the case now), recognizing that some of the contradictory results may simply reflect the quality of the data rather than an inadequacy of the theory. In evaluating the theory, we should expect efforts to think through the measurement of the dependent variable, and especially important is a willingness—where necessary—to examine the consequences of the theory with the outcomes observed in some situations, regardless of how tightly the causal mechanisms have been measured. Ultimately, this model is a communal model—it depends on

the efforts of a community of researchers pursuing different facets of the same problem. We remind the reader that this is precisely the guidance from Fisher reported by Cochran in the quotation cited in Section 5 of this paper. Ultimately, the outcome will depend on the ability to weigh and evaluate the body of evidence to reach a conclusion about the theory. But such an evaluation encompasses the work of many.

7. EXAMINING ALTERNATIVE THEORIES AND EXPLANATIONS

If an initial theory yields reasonably promising results, then the next step in Implication Analysis is to list and evaluate alternative explanations "including all different hypotheses and biases in the results" (Cochran 1965:252–53). By bombarding the initial theory with alternatives, researchers ask if there is an alternative that can not only account for the same observations as the first one, but does it also explain additional observations? Or, even if an alternative theory fails to account for some observations covered by the first, does it cover events for which the initial theory cannot? Although we would like to deal with an alternative of at least comparable scope, also relevant are narrower causal propositions—not as broad in magnitude and range as the initial one—if they reveal a serious weakness in the initial theory. Finally, there is always the challenging question of whether the initial causal connection is illusory, reflecting a spurious presence between the initially hypothesized cause and the "true" causal force (for example, see the discussion of arsenic and lung cancer below).

At first glance, this second phase may appear to be nothing more than intellectual masochism. Since it is hard enough to find a theory that will hold up for a broad range of events, why go out of the way to consider if a "successful" theory can be superseded or at least needs modification? Given our earlier review of the obstacles that the social sciences usually face, we cannot be satisfied with a result that is simply harmonious with the causal theory. Rather, it is necessary to ascertain if other theories and possible biases can also account for these results. Until then, the initial results should be described only as *promising*.

It is unlikely that the available evidence—if subjected to a rigorous examination—will fully support the initial theory on all counts.

This is not a crucial problem if we take a probabilistic rather than a simple dichotomous view of a theory that the evidence says is *true* or *false*. Given the nature of the data as well as the difficulty in coping with a complex environment all at once, not all of the results are likely to work out as expected—even if the theory is extraordinarily robust. Of value, on this score, is Popper's distinction between what he calls the "pre-scientific" and "scientific" levels of evaluating the soundness of a theory. The scientific standard, Popper explains, "presupposes that we can look at our theories critically—as something outside ourselves. They are not any longer our subjective beliefs—they are our objective conjectures" (Magee 1971:73). This can be seen as Popper's way of distinguishing between conclusions that are reached after an extensive and rigorous consideration of the theory and the evidence, as opposed to conclusions that are reached when the investigator finds *some* evidence in support of a theoretical claim and then chooses to use these results to mount an argument in its favor—even when the evidence is far from being as extensive or as conclusive as can be generated in the social sciences. (This is far more beneficial, in our judgment, than his misunderstood "falsification" idea.) It is simply *necessary* to address alternatives as a way of ultimately reaching some consensus about the superiority of one or another theory, or quite possibly consider whether a need exists for modification or development of a new alternative. Some examples are presented below of the process through which an established theory is challenged by alternative causal propositions.

7.1. *Lung Cancer*

At first glance, this example dealing with lung cancer and the following one on cholera may appear inappropriate for a discussion dealing with social research. They are included for a special reason. In both cases, the initial research and conclusions were based on the same type of data that we typically encounter in social science: no experimentation and initially somewhat vague theories accompanied by the need to weigh alternative theories and speculations. Both efforts are later supported by the kind of evidence all too rarely seen in sociology. In that sense, we can see a powerful validation of the types of procedures that can be followed in the social sciences in the absence of experimentation and accompanied by potentially damaging questions of selectivity. (Fisher himself asserted

that people who smoked were more likely to get lung cancer because they were *initially different* from nonsmokers—as opposed to getting lung cancer as a *consequence of smoking*.)

The review by Doll and Hill (1952) of the reported association between smoking and lung cancer is cited by Cochran (1965) as an elegant example of the reasoning and evidence used by imaginative researchers to evaluate alternatives to the theory attributing the smoking of cigarettes as a major cause of lung cancer. Admirable as their groundbreaking work may have been at its time, Doll and Hill's procedures are relevant for present-day social science. In fact, their reasoning is highly applicable to some of the major problems addressed today. As the biological mechanism linking the ingredients in tobacco with lung cancer was not yet known and there was no experimental evidence to consider, their review was based largely on observational data of the sort that the social scientist is still often obliged to use. (Keep in mind that this paper addresses not the continuous advances in statistics but rather ways of thinking about the evidence—whether it is statistical, experimental, or qualitative.)

Arsenic provided a challenging causal alternative to the hypothesis that tobacco is the causal ingredient responsible for the higher frequency of lung cancer among cigarette smokers. Since arsenic was a widely used ingredient in the insecticides sprayed on tobacco plants, it was plausible that arsenic—rather than the tobacco—was the source of the lung cancer frequently occurring among smokers. (At that time, the agent[s] in tobacco smoke responsible for cancer had not been located.) This is not a trivial alternative since a radically different public health solution would be to eliminate the arsenic in insecticides rather than discouraging smoking *per se*. Finding that arsenic was not used for growing tobacco in Turkey, Doll and Hill (1952) explored the cancer rates in that nation. They found that the frequency of lung cancer in Turkey was in line with the level of smoking in that country, when compared with rates for other nations where tobacco plants were exposed to arsenic—hardly supporting this alternative interpretation. Note how the determination of the mechanism is also crucial for evaluating causal relations in social research—particularly when the investigations are intended for policy applications.

A variety of alternative explanations were evaluated by Doll and Hill through the use of appropriate evidence. For example, they even considered the use of petrol lighters by smokers as the causal factor—not

the smoking *per se*—and found evidence to reject this. Since only smokers would be using cigarette lighters, how did they reach this conclusion? They compared the lung cancer rate of smokers who used lighters with smokers who used matches thereby providing an estimate of the influence of lighters *per se* on lung cancer rates. Since lung cancer had been reported as especially common among gasworkers, they considered that factor and found it wanting. They also found the "use of coal, gas, or electric fires or other forms of heating in the living-rooms of their homes did not differ appreciably" between patients with lung cancer in comparison with patients suffering from other diseases (p. 1284). They evaluated possible problems in the quality of the data. For instance, they considered whether the information on smoking behavior collected from patients with lung cancer was warped such that "the lung-carcinoma patients tended, because of their disease, to exaggerate their smoking habits" or "that the interviewers tended to scale up the smoking habits of the lung-carcinoma patients" (p. 1282). But the possibility of an interviewer effect was also ruled out. They determined that the smoking rates compiled by interviewers for patients initially diagnosed as having lung cancer (but later found to have a different disease) were comparable to other non–lung cancer patients and much lower than those compiled for lung cancer patients. Considerable attention was also paid to the possibility that social class played a role in lung cancer for a variety of reasons such as occupational differences in the exposure to carcinogenic risks. They found that class *per se* was not a factor at all.

Doll and Hill (1952) also effectively eliminated various hereditary possibilities by considering the shifts over time in smoking and the corresponding shifts in lung cancer (hardly likely to reflect a sudden genetic shift) and the lag in lung cancer for women, when compared to men, was found to be closely linked to the lag in smoking. Finally, they were able to evaluate and provisionally reject the alternative theory that people with a certain constitution were more likely to smoke and therefore more likely to develop lung cancer—in other words, that the association between smoking and lung cancer was spurious, not causal. The grounds for rejecting this proposition are of interest because they illustrate that the reasoning behind certain theories can be evaluated by considering some basic data along with the implication of the theory. In this case, if lung cancer is indeed not caused by smoking but by a physical condition that leads those prone to develop lung cancer to also

smoke, then Doll and Hill show that it would take some very strong alternative environmental factor to account for the enormous growth in lung cancer at the same time that there is a concomitant growth in smoking. Put another way, lung cancer was relatively rare prior to the growth in cigarette smoking, but the physical condition leading to the development of lung cancer outside of smoking presumably would have been present prior to smoking's popularity.

Many of these observational steps are of special interest because inferences are not based on direct evidence. When we consider the growing disposition to attribute social outcomes to genetic causes, the work of Doll and Hill (1952) is an even more important model. But we should not lose sight of the main goal—addressing and evaluating alternative causal explanations. That is really the case for attributing the occurrence of lung cancer to biological sources such as the propensity of those disposed to lung cancer to also be the ones who smoke—a reversal of the causal direction in the initial theory under consideration. Instead, the alternative hypothesis is evaluated by considering its implications and then ascertaining whether the expected indirect evidence is consistent with the alternative theory.

In the absence of suitable direct evidence, note that imaginative steps can be taken to evaluate a theory in terms of its observable consequences. For example, sociobiologists attribute various facets of social behavior to genetic sources. It is very difficult—at least at this point—to directly evaluate such claims. But it is another matter to evaluate the implication of the theory. If the attraction between the sexes were genetically driven in terms of survival of the species, this would imply relative constancy of the feature's appeal over time as well as a direct correlation between the physical feature and its benefit for survival and reproduction (allowing for historical contexts). The failure of these implications to hold would obviously not be a definitive negation of the theory, but it would require even stronger evidence in support of the sociobiological theory—at least with respect to attractiveness between the sexes.

7.2. *Cholera*

The causal analysis of cholera by pioneering epidemiologist John Snow provides another superb example of how an established theory and an alternative can be evaluated by weighing the empirical implications of

one against the other. Note that this is not in the simplistic form of using some specific case of a predicted outcome as "a crucial test" of the two theories. Here, observations of the differences in the causal processes and mechanisms implied by the two theories are used to evaluate the competing theories. Our entire discussion is drawn from Freedman's excellent review of John Snow's work on cholera (1993). The dominant theory in the 1850s was that disease was caused by minute, inanimate poisonous particles in the air, called "miasmas." Freedman reminds us that the microscopes of that time were of such low quality that human pathogens could not be seen and the infection theory had a limited number of supporters.

Snow observed that cholera had a history of striking in waves, with early symptoms being vomiting and diarrhea. This led him to speculate that there was a living organism that entered the body with food or drink, multiplied in the body, and then passed out and got into the water supply, thereby infecting new victims. Snow's task consisted not only of developing his theory to account for cholera but also evaluating the dominant alternative explanation in terms of how well it could account for the epidemiological features of cholera. He noticed that cholera has a specific pattern of spreading. It tracks human commerce. "If a ship goes from a cholera-free country to a cholera-stricken port, the sailors get the disease only after they land or take on supplies. The disease strikes hardest at the poor, who live in the most crowded housing with the worst hygiene. These facts are consistent with the infection theory and hard to explain with the miasma theory" (Freedman 1993:295).

Snow used all sorts of observational data in a continuous interaction between the evidence and his theory. He found that the first case of a new epidemic was attributed to a seaman, newly arrived in London from Hamburg—where the disease was prevalent. He found the second case was a man who had taken a room in the place where the first carrier was living. He observed that the residents in one of two adjacent apartment buildings were badly hit by cholera, but not the residents in the neighboring building. In turn, he found that the water supply for the first building was contaminated by runoff from privies whereas the water in the second building was substantially cleaner.

The cholera epidemic occurred during a period in which there were various private companies supplying water in the same areas of the city. Some took their water from polluted sources, such as the

Thames, while other companies used cleaner sources. A natural experiment, therefore, was possible because the water companies served many of the same neighborhoods—indeed some were serving houses on the same block. Given that the consumption of clean versus polluted water in a block was not a function of other characteristics of the residents, the linkage between the water source and the incidence of cholera could be attributed to the water supply itself. This, in turn identified a causal connection between the cleanliness of the water supply and cholera. Freedman (1993:295) goes on to show how Snow's spot maps during the 1853–1854 epidemic in London provided strong support for the causal connection between the quality of the water supply and cholera, with the incidence of cholera piling up around the pumps supplying contaminated water.

There are other examples of how Snow evaluates the implications of his theory, which miasma theory cannot explain. For example, in a location where there was a high incidence of cholera around the pump supplied by contaminated water, he found a poorhouse with very few cases. He discovered that the poorhouse had its own well and the inmates did not take water from the contaminated pump. He also found a brewery with no cases of cholera; it was one where the workers drank beer, not water. Bear in mind the crucial ecological correlation between the levels of cholera among residents connected to a water company drawing water above the main sewage discharge points in the Thames, as contrasted with comparable residents using water from companies that used intake points downstream from the sewage discharge points. A table constructed by Snow, based on the experiences of 300,000 people, shows a remarkably higher level of deaths among those residing in houses drawing their water from polluted sources. (See Freeman's discussion beginning on p. 296, which includes on p. 298 a summary of Snow's table.)

Other events supported his conclusions. New York City unsuccessfully treated the cholera epidemics in 1832 and 1849 with such methods as calling for temperance from the residents; bringing in pure water to *wash the streets*; and treating the sick by bleeding and mercury. When the implications of Snow's theory were followed in the 1886 epidemic (boiling the drinking water, isolating the sick, and disinfecting their bodily wastes), the death rate was less than one-tenth of the earlier rates. (Freedman also traces the later biological evidence that fully supported Snow's theory.)

7.3. *Durkheim's Analysis of Suicide*

Because the pioneering study of suicide by Emile Durkheim (1951) is well known, we will only note that much of the rich empirical work in that volume can be seen as harmonious with the suggestions going back to Cochran's approach inspired by Fisher, and that it is also consistent with the response to inconsistent evidence that we will discuss next. Namely, the various theories of suicide that others had advanced are bounced against the evidence that Durkheim gathers to evaluate them. This is in turn followed by alternate theories of suicide proposed by Durkheim, who then weighs them with the available evidence. Whether the reader wishes to consider this as an exemplar of social science research is not the issue (witness the heavy reliance on what we now call "ecological correlations"). Rather it is an extraordinarily early example of the dialogue between data and competing theories; the evaluation of various theories is almost entirely based on observational data that are less than ideal. He addresses each theory—whether it is one that he eventually supports or rejects—with all sorts of data meant to get at the implications of the theory as best as is possible. More about this later, when we deal with an almost inevitable outcome of this research proposal—to wit, the results will not be entirely consistent.

8. EVALUATING INCONSISTENT EVIDENCE: JURY TRIALS AND CHECKLISTS

For good reason, multiple empirical evaluations of a theory are apt to generate inconsistent findings that are difficult to interpret. Is there a problem with the theory? Or does the inconsistency reflect the data used to evaluate the theory? Or is there a more complicated situation such that the theory is valid but other conditions need to be considered? Sound theories might very well stumble because of data problems. By the same token, inadequate or inappropriate evidence may appear to support unsound theories. Of course, we cannot ignore largely consistent results in one direction but even then we have to proceed with caution. Consistent empirical support of a theory is a good sign but is not necessarily definitive. Likewise, an occasional contradictory result need not mean that the theory is invalid or in need of modification. One way or another,

it is necessary to consider if inconsistent results are reflecting a problem with the data or a weakness in the theory.

8.1. *Resolving Inconsistent Evidence*

Since the analysis of inconsistent and incomplete evidence is itself underdeveloped, our observations in this section apply to various research "strategies." Inconsistent evidence can rarely be resolved with the procedures typically employed by social science. Theory testing, tests of significance, principles of parsimony, estimates of predictive power, variance explained, usage of control variables, simulation of true experiments, and other routine procedures are unlikely to accomplish this task. This is because the contradictions must consider the quality and relevance of the evidence. As such, our discussion should be viewed as what we hope will prove to be a starting point for a more extensive way to approach results that appear to largely, but not fully, support one theory. This occurs in two interesting forms: (1) when not all of the results can be explained by the most promising theory, but no other theory can successfully encompass the remaining results; and (2) when neither of the theories can account for essentially all of the results, but together they provide a fairly complete accounting. (There are other more complex difficulties that should be addressed, but this is enough for a starter.)

In short, our usual ways of dealing with inconsistent or contradictory results are predicated on closing our eyes to the situation that the social researcher typically encounters. The disposition is understandable: go forward with what one has. A case is made for this since—hopefully—something is better than nothing and being hypercritical without decent alternatives is nonproductive when we find results that the researcher would like to support. Why look further? In particular, why run the "risk" of finding a different result? In no small way, the shortcomings we encounter can be overcome with the alternatives described below.

8.2. *A Jury Trial*

The jury trial is an appropriate and useful analogy for the social sciences. Just as all of the evidence in the trial may not consistently point in

the same direction, likewise we can expect that not all of the research evidence will point in the same direction when a theory is evaluated. To illustrate, let's imagine ourselves as members of a jury. The defendant has been charged with breaking into a bank and stealing a stash of hundred-dollar bills. (Think of this as a causal theory.) In support of this charge, the prosecutor presents evidence. The prosecutor calls on eyewitnesses, fingerprint experts, the police who found the money in the defendant's car trunk, and the like. The defense in turn presents evidence to show us that the defendant did not break into the bank and that there is instead an alternative explanation of the prosecution's evidence (in other words, that the causal hypothesis is not correct). The defense provides testimony showing that the defendant was elsewhere at the time of the crime; a demonstration that the lighting and distance between the accused and a witness would make definitive identification impossible; and an alternative explanation for the fingerprints and the banknotes in the car (revealing that the defendant found the money on the street and was actually en route to a police station when stopped). Of course, this analogy is a bit of a stretch since hopefully all researchers are seeking the truth without resorting to rhetoric and other devices meant to hide all the relevant facts. (In a criminal case, the defendant does not want a lawyer who shows that there are six reasons to think the client is innocent, albeit pointing out three reasons in the opposite direction.)

Note that the failure of one part of the prosecution's case or the defense's case does not solve the issue. If the defense shows that the distance and lighting conditions reduce confidence in the eyewitness identification of the defendant, this is not necessarily a crucial test—after all, there is other evidence to support their case than just this. Even a persistent negative result in a complex case will not necessarily mean the theory is "wrong," and instead it shows only that the theory is not supported by one particular facet of the evidence.

A crucial factor is not the result *per se*, but what the result contributes to our evaluation of the theory. Evidence for or against a theory is good, but there are two other questions that have to be addressed. One is what we can call a *technical question*—namely, given the preponderance of evidence pointing in one direction, can we account for the minority of results in the opposite direction? Note that a probabilistic perspective does not require such an accounting. But as we move from situations of overwhelming evidence in one direction to a more difficult

result where the predominant outcome clashes with a modest number of contrary results, then we will be happier if we can resolve this. The second question to consider is whether the result is a crucial one. In the jury trial, for example, the evidence introduced by the defense about lighting and distance has undercut our confidence in the eyewitness identification of the defendant. Does this matter? There is no simple answer because the answer depends on whether the identification was crucial. Is there a sufficient body to convict—i.e., to support the theory? Or does this wipe out the heart of the case?

8.3. *Technical Issues: A Basic Checklist*

Data quality is a dynamic issue; inevitably, what is considered suitable data will change over time as knowledge about a topic proceeds. The first step in evaluating contradictory or inconsistent results is a prosaic one. Rather than considering new theories or modifying a theory to exclude the deviant cases (for example, see the use of scope statements proposed by Walker and Cohen [1985]), or looking for neglected control variables that might explain the difficulties, we must ask about the quality and nature of the data. In essence, we need a checklist of possible difficulties due to a variety of differences in the procedures, measurements, context, sampling, and the like that could account for inconsistent or contradictory results. How good are the data? How well do the data measure what we want to know? We know the usual answer: *not as good as we would like, but the best we can work with—given the limits of money, time, place, or practicality*. Fair enough, but this does not mean that the issue should be ignored, particularly when evaluating conflicting results. For instance, if some of the results go one way and some go another, are the results associated with the different types of data or with some other procedure?

This checklist requires serious consideration about such matters as the form of the question used in a survey; the sample; the coding of the data; possible biases of the organizations conducting the survey; distortions due to the interviewers or researchers; and "adjustments of the data." We have even witnessed dichotomous data that were coded incorrectly, so that what appeared to be an association in one direction was in fact occurring in the opposite direction. No type of data is immune from these matters: Just as studies provided by commercial or policy

organizations are susceptible to errors, so too are official governmental data. Likewise, surveys and ethnographic studies each have their problems. Only recently, for example, has there been a greater focus on the need to monitor and minimize the danger of unintentional distortions causing inconsistencies on respondents' answers (see Yin [2003]). And it is appropriate to ask this question of other data as well—for example, inconsistent results between "hard" and "soft" data.

Incomplete responses are another difficult problem that can generate inconsistent results. Some surveyors go to great lengths to increase the response rate or to estimate the distortions resulting from incomplete responses. There has been some progress since Wiley (1984) illuminated the distortions in conclusions that can result from nonresponse in surveys—whether they are respondents who initially refuse to answer a survey or who are lost in the second wave of a survey. Recently, we have found sound evidence that nonresponse need not always mean selectivity leading to a serious distortion, but certainly it does operate as a major consideration in evaluating the results. (Singer [2006] includes a diverse set of results about the connection between nonresponse rates and nonresponse bias.) In any case, we must consider nonresponse bias as a possible source for inconsistent results.

The statistical techniques applied to data analysis are probably the most common issue considered in discussions about results generally, let alone the difference in results. It is a specialized matter that merits—and receives—extensive discussion. In passing, we note its relevance for the checklist of technical issues. The assumptions underlying many statistical procedures are often neglected or ignored, with dire consequences for the conclusions obtained. In addition to the question of what is the *best* method, there is also the question of whether the choice of method accounts for the inconsistency or consistency in the outcomes observed.

There is another deep issue that is so obvious and pervasive and seemingly intractable that it tends to be ignored. Do the data measure what we want to measure? More subtly, we want to know how closely the actual data approximates the ideal data. This is tricky to contemplate. We know, of course, that rarely do we have ideal data, and that it can be a pointless exercise to deal with a task that cannot be fully mastered or even measured very well. On the other hand, the gap between the *actual* and the *ideal* can easily affect consistency in the results and interpretation of the empirical outcome—whether it be largely consistent

with the theory, largely inconsistent, or somewhere in between. This is a particularly important matter when working with observational data not initially gathered or developed by the investigator, but it is of equal importance when the researcher has gathered the data—whether it is observational or even experimental. There is a strong chance that data developed by some other researcher or organization are not exactly the measure that the researcher would have wanted to use. Not all discrepancies are equal, as it were, and therefore some studies of the same topic may be closer to the ideal than others (the *ideal* being what would have been chosen from a universe of all possible data sets).

Even with experiments based on random assignments, the researcher may end up with data that are not exactly the *ideal*—in the context described above. After all, there are limits on what researchers can do to people, or even ask them, let alone limits on the availability of the information for organizations, nations, and the like. There can be improvements on surrogate measures, but surrogate measures are still surrogate measures. Therefore, it is always necessary to consider the appropriateness of the data. First, if the results are inconsistent, does this merely reflect the differences in the nature of the data and its quality? Second, regardless of consistency issues, do the data provide a reasonable approximation of what would be the ideal measure? This is a serious problem in counterfactual analyses since extensive statistical manipulation will be required and the measurements are more likely to deviate from the ideal. Since the typical counterfactual analysis uses data that differ from the ideal, this means that the results are counterfactual statements about counterfactual estimates.

We can think of theories as varying along a continuum from a narrow range to a broad range of empirical ramifications. This generates another technical problem. As a general rule, the obstacles to evaluating a theory properly vary directly with its breadth. A theory with a narrow range of applications is relatively easier to evaluate because the ramifications of the theory entail less variation in the contexts that need to be considered. For a valid theory of this type, inconsistent results are less probable (assuming confidence in the data) and contradictory results are more likely to signal rejection or the need to modify the theory. The downside is that the theory has less range.

In contrast, a theory that covers a wide range of situations will require far broader empirical examination and, as such, will encounter increasingly heterogeneous situations involving different

context variables, measurement problems, and the like. We can expect more a priori situations where the theory will fail to hold—not because it is "wrong" but simply because there will be the need to take into account other circumstances that could overpower the expectations generated by the theory. Of course, the greater diversity of situations means more and more opportunities for errors and problems in the basic measurement issues. This leads us to a straightforward proposition: The standards for evaluating a theory should take into account the nature of the theory. As a wider variety of contexts are encountered, there is a greater chance that the implications of the theory under consideration will be masked by the influences due to other factors that get in the way. As the range of a theory increases, the likelihood of finding inconsistent evidence will increase—even if it is valid—because the array of variables to be included will also increase. As the variety of attributes (considered in at least one setting) increases, there is more exposure to measurement error. The importance of reviewing the various technical issues is all the more vital.[5]

9. CONCLUSIONS

The quality of evidence in the social sciences rarely allows for the evaluation of a theory in a manner comparable to what we believe occurs in the hard sciences. Nothing is gained by acting as if the conventional data and standards in the social sciences permit strong conclusions. Fortunately, more can be done than simply wringing our hands and bemoaning the absence of "science-like" substitutes. There is no point in retracing all of the steps described in this paper, but we remind the reader of several key points.

First, there is a distinction between finding a specific result that is consistent with a theory as opposed to a body of results (incorporating different contexts and different alternatives) that supports the theory.

[5] Although Walker and Cohen (1985) have also noted the increasing likelihood that evidence for broad theories may often be contradictory, we quickly part company in the analysis of both the cause and in their solution. They take a deterministic view of theories such that an exception or an inconsistency is deeply damaging to a theory, requiring development of scope statements in order to modify the theory and thereby "save" it. Quality of evidence is not an issue for them. The "failure" of a theory is never attributed to the series of obstacles we describe.

Consider an example drawn from Aczel (1997:103). Suppose I observe seven men and seven women in a room. I conjecture that these are seven married couples. My companion responds that while there is no obvious reason for rejecting the statement as false there is no reason for accepting the statement as true either. The presence of seven men and seven women in the room hardly rules out a variety of other possibilities. Just to name a few: all of the people could be single; some could be married, others not; they could all be part of a large family (brothers, sisters, offspring, aunts, uncles, cousins); they could be coworkers. Obviously, more empirical work is needed before we could be satisfied with the correctness of the conclusion. The fact that the married-couple hypothesis is consistent with the limited evidence does not mean that it is true. The accumulation of data in favor of this hypothesis and inconsistent with rival hypotheses is necessary in order to increase our confidence. In a more general way, we can say that the result for the "seven married couples" is a *dangling theory*. It may or may not be true, but the evidence is insufficient. We must also recognize that the inadequacy of the data tells us nothing about the adequacy of the theory. It means only that we do not know enough with the available evidence. Unless it is truly impossible to obtain the necessary evidence now (or in the future), inadequate data are not grounds for either rejecting or accepting a dangling theory. To do so is to commit a very common mistake in the social sciences: concluding that the absence of information means that the theory (or conclusion) under consideration is false.

Second, a large number of observations and the consideration of both a variety of contexts and alternative theories are required to evaluate a theory. This is particularly the case for an important and broad theory. Our technique is not widely practiced, nor is it novel. Witness, for instance, Griswold's (2001) observation regarding studies "that accumulate so much evidence and subject it to such painstaking scrutiny that one cannot help but be convinced by them; the thesis is not so much argued as established" (p. 1828). In addition, the results need not be due to the work of one person or research group, but more likely will be a communal effort, as was the case for the studies of smoking (although obviously some were more prominent than others). In any event, these smoking studies would meet Griswold's observation of "a careful but relentless gathering and weighing of the evidence" (p. 1828). In similar fashion, Rosenbaum (2002) notes that "Scientific questions are not settled on a particular date by a single event, nor are they settled

irrevocably. We speak of the weight of evidence. Eventually, the weight is such that critics can no longer lift it, or are too weary to try. Overwhelming evidence is evidence that overwhelms responsible critics" (p. 11).

Bear in mind that the goal is not the "proving" of a theory, but something more like the "strong support" of a theory when there is a large body of evidence in favor and alternative explanations are very implausible. On this score, it is important to view our theories as probabilistic since we do not live in an ideal world and are often obliged to use less than ideal evidence. Theories will not always work out—even if they are "correct." This is implicitly recognized by Popper (2002), who observes that "non-reproducible single occurrences are of no significance to science. Thus a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified" (p. 66). The widely held and warped image of what Popper meant by *falsification*—to wit, an unforgiving deterministic standard operating with precisely measured data—if actually followed, would sooner or later conclude that every theory is false. As a consequence, those who still adhere to this position (hopefully, a small and declining number) are enmeshed in an unreasonable and unacceptable standard that accomplishes nothing but harm. Even in the absence of errors in data, we can be confident that all theories can be shown to be false—simply because it is impossible to specify all possible conditions and, therefore, a literal interpretation of what a theory implies can be taken out of context and lead to a negative result (Lieberson 1998).

Third, the boundaries drawn between quantitative and qualitative analysis, as well as the boundaries between different methods more generally, do not justify the automatic denial of the potential value of any method. There is a distinction between the type of procedures that a researcher may favor and the consideration of data from all sources. If we are to generate as many implications of both a theory and its alternatives, then this means gathering as much data as possible to address the theoretical issues. Data are then tightly evaluated to help us draw a sound conclusion. The quality of the data becomes highly relevant as it is almost certain that there will be inconsistencies in the evidence. We are not naıve enough to think that we can expect researchers and theorists to work on all fronts, since they have their skills and dispositions, but the key point is that *all of the results should be incorporated and considered in the evaluation of the theory*. The distinction between a qualitative theory of some phenomena and a quantitative theory of

the same phenomena is artificial and nonproductive. If, however, different types of data persistently give different conclusions, then it is an appropriate challenge to think through the source of this inconsistency.

Finally, just as the initial efforts to evaluate a causal hypothesis or a theory can be viewed as a work in progress, so too should our analysis of implications. Several developments will help improve its utility: (1) the expansion and elaboration of the checklist we laid out in this paper. Undoubtedly, there are many other technical issues to consider when contradictions appear to exist. (2) We have to recognize that the social sciences operate with a distorted image of practices in the hard sciences. There is the tendency to assume that the quality of the evidence solely reflects the quality of the theory. If a theory is valid, there is an implicit expectation that one study (or at most, a handful) will yield evidence that is precise, crisp, and supportive of a parsimonious theory. At best, this will rarely occur and our evaluation of the theory is likely to be lowered because of it. The evidence certainly should reflect the theory, but it might also reflect unrelated circumstances—namely, the difficulty in obtaining the precise and accurate evidence that we assume should occur for a robust theory. Given the empirical obstacles in the social sciences, this can be an inappropriate expectation. Almost certainly there are other implicit assumptions about the linkage between theory and evidence that Implication Analysis needs to examine and root out. (3) We must adjust our standards by moving away from an *exclusive reliance* on mechanical ways of evaluating evidence such as tests of significance, predictive power; and variance explained. In our estimation, the jury trial model discussed earlier may well prove to be a more appropriate way of proceeding in the social sciences. This means evaluating a theory by using data from diverse sources, of different quality, and meeting different levels of confidence in their accuracy and relevance, and they are possibly somewhat contradictory. It is more a matter of weighing the evidence by piecing together a complex array of information. There is no simplistic rule for doing this. (See for example Ni Bhrolchain and Dyson's [2007:29] use of the metaphor *bricolage* as "knitting together diverse strands of evidence.") Certainly, experiments (true ones or those based on observational data) would be important to consider. But the key step is that the evaluation is based on all of the evidence—albeit not equally. Note how this is very different from using experiments (real, statistical, or natural) as the information that trumps all other information. There is an inherent tension: On the one hand, formal mechanical

methods and standards convey a sense of objectivity and rigor: On the other hand, a formal set of rules cannot replace judgment and good sense that should be of help in evaluating the appropriateness of the evidence. The latter is hardly trivial, when one considers the reliance on data that is rarely of the form that we would ideally prefer.

In the same way that the hard sciences evolve and develop through the efforts of worker bees, so too should the social sciences recognize that worker bees are adding to a well-structured problem and the outcome is the product of all of the work. Great people exist, but their theories, as well as their syntheses of the existing work, are open to modification and are neither proven nor disproven with a given data set.

REFERENCES

Aczel, Amir D. 1997. *Fermat's Last Theorem*. New York: Bantam Doubleday Dell Publishing Group.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgment of Taste*. Cambridge, MA: Harvard University Press.

Brush, Stephen G. 2003. "'Scientific Method': Guide to Research, Principal of Demarcation, or Neither." For the Berkeley Conference on Utility of History of Science. Berkeley, CA.

Cao, Yang, and Victor G. Nee. 2000. "Comment: Controversies and Evidence in the Market Transition Debate." *American Journal of Sociology* 105(4):1175–89.

Cochran, William G. 1965. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society* 128(2):234–66.

Cohen, Moris R. 1934. *An Introduction to Logic and Scientific Method.* London: Routledge and Kegan Paul.

Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of National Cancer Institute* 22:173–203.

Doll, Richard, and A. Bradford Hill. 1952. "A Study of the Aetiology of Carcinoma of the Lung." *British Medical Journal* 2:1271–86.

Durkheim, Emile. 1951. *Suicide*. Glencoe, IL: Free Press.

Erickson, Bonnie H. 1996. "Culture, Class, and Connections." *American Journal of Sociology* 102(1):217–51.

Firebaugh, Glenn. 2007. "Replication Data Sets and Favored-Hypothesis Bias: Comment on Jeremy Freese (2007) and Gary King (2007)." *Sociological Methods and Research* 36(2):200–9.

Freedman, David A. 1993. "Statistical Models and Shoe Leather." Pp. 291–313 in *Sociological Methodology*, edited by Peter. V. Marsden. Oxford, England: Blackwell Publishing.

Freese, Jeremy. 2007a. "Overcoming Objections to Open-Source Social Science." *Sociological Methods and Research* 36(2):220–26.

———. 2007b. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods and Research* 36(2):153–62.

Gerber, Theodore P., and Michael Hout. 1998. "More Shock Than Therapy: Market Transition, Employment, and Income in Russia, 1991–1995." *American Journal of Sociology* 104(1):1–50.

Griswold, Wendy. 2001. Review of *A Matter of Taste: How Names, Fashion, and Culture Change* by Stanley Lieberson. *American Journal of Sociology* 106(6):1826–28.

Halle, David. 1993. *Inside Culture: Art and Class in the American Home*. Chicago: University of Chicago Press.

Harrington, David P. 2000. "The Randomized Clinical Trial." *Journal of the American Statistical Association* 95(449):312–15.

Hill, Austin Bradford. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58(6):289–300.

Holton, Gerald. 2004. "Intuition in Scientific Research." Department of Physics, Harvard University. Unpublished manuscripts.

King, Gary. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods and Research* 36(2):173–99.

Lamont, Michele. 1992. *Money, Morals, and Manners: The Culture of the French and the American Upper-Middle Class*. Morality and Society series, edited by Alan Wolfe. Chicago: University of Chicago Press.

Lieberson, Stanley. 1963. "The Old-New Distinction and Immigrants in Australia." *American Sociological Review* 28(4):550–65.

———. 1980. *A Piece of the Pie*. Berkeley: University of California Press.

———. 1992. "Einstein, Renoir and Greeley: Some Thoughts About Evidence in Sociology." *American Sociological Review* 57:1–15.

———. 1998. "Examples, Submerged Statements, and the Neglected Application of Philosophy to Social Theory." Pp. 177–91 in *What Is Social Theory? The Philosophical Debates*, edited by Alan Sica. Malden, MA: Blackwell Publishers.

———. 2000. *A Matter of Taste: How Names, Fashion, and Culture Change*. New Haven: Yale University Press.

Lieberson, Stanley, and Freda B. Lynn. 2002. "Barking Up the Wrong Branch: Scientific Alternatives to the Current Model of Sociological Science." *Annual Review of Sociology* 28:1–19.

Magee, Bryan. 1971. "Conversation with Karl Popper." Pp. 66–82 in *Modern British Philosophy*. London: Secker and Warburg.

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

McCloskey, Deirdre N. 1998. *The Rhetoric of Economics*. Madison, WI: University of Wisconsin Press.

Mill, John Stuart. 1874. *System of Logic*. New York: Harper.

Moffitt, Robert A. 2004. "The Role of Randomized Field Trials in Social Science Research." *American Behavioral Scientist* 47(5):506–40.

———. 2005. "Remarks on the Analysis of Causal Relationships in Population Research." *Demography* 42(1):91–108.
Nee, Victor. 1989. "A Theory of Market Transition: From Redistribution to Markets in State Socialism." *American Sociological Review* 54(5):663–81.
———. 1991. "Social Inequalities in Reforming State Socialism: Between Redistribution and Markets in China." *American Sociological Review* 56(3): 267–82.
———. 1996. "The Emergence of a Market Society: Changing Mechanisms of Stratification in China." *American Journal of Sociology* 101(4):908–49.
Nee, Victor, and Yang Cao. 1999. "Path Dependent Societal Transformation: Stratification in Hybrid Mixed Economies." *Theory and Society* 28(6):799–834.
Nee, Victor, and Rebecca Matthews. 1996. "Market Transition and Societal Transformation in Reforming State Socialism." *Annual Review of Sociology* 22: 401–35.
Ni Bhrolchain, Maíre, and Tim Dyson. 2007. "On Causation in Demography: Issues and Illustrations." *Population and Development Review* 33(1):1–36.
Nowak, Rachel. 1994. "Problems in Clinical Trials Go Far Beyond Misconduct." *Science* 264(June 10, 1994):1538–41.
Popper, Karl. 1964. *The Poverty of Historicism*. New York: Harper.
———. 2002. *The Logic of Scientific Discovery*. New York: Routledge Classics.
Quine, W. V., and J. S. Ullian. 1978. *The Web of Belief.* 2nd ed. New York: McGraw-Hill.
Rogowski, Ronald. 2004. "How Inference in the Social (but Not the Physical) Sciences Neglects Theoretical Anomaly." Pp. 75–83 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by Henry E. Brady and David Collier. Lanham, MD: Rowman and Littlefield.
Rosenbaum, Paul R. 2002. *Observational Studies,* 2nd ed. New York: Springer-Verlag.
Ross, Edward Alsworth. 1914. *The Old World in the New*. New York: Century Company.
Schliesser, Eric. 2005. "Galilean Reflections on Milton Friedman's 'Methodology of Positive Economics,' with Thoughts on Vernon Smith's 'Economics in the Laboratory.' " *Philosophy of the Social Sciences* 35(1):50–74.
Seawright, Jason, and David Collier. 2004. "Glossary." Pp. 289, 304 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by Henry E. Brady and David Collier. Lanham, MD: Rowman and Littlefield.
Singer, Eleanor, ed. 2006. "Nonresponse Bias in Household Surveys." *Public Opinion Quarterly Special Issue* 70:637–45.
Smith, Vernon L. 1994. "Economics in the Laboratory." *Journal of Economic Perspectives* 8(1):113–31.
———. 2002. "Method in Experiment: Rhetoric and Reality." *Experimental Economics* 5:91–110.
Szelenyi, Ivan, and Eric Kostello. 1996. "The Market Transition Debate: Toward a Synthesis?" *American Journal of Sociology* 101(4):1082–96.

Walker, Henry A., and Bernard P. Cohen. 1985. "Scope Statements: Imperatives for Evaluating Theory." *American Sociological Review* 50(June):288–301.

Wiley, James A. 1984. "Destructive Testing of Survey Findings." Survey Research Center, Univeristy of California, Berkeley. Unpublished manuscript.

Winship, Christopher. 2007. "Editorial: Introduction to the Special Section on Replication and Data Access." *Sociological Methods and Research* 36(2): 151–52.

Yin, Robert K. 2003. *Case Study Research. Vol. 5, Design and Methods*. Thousand Oaks, CA: Sage Publications.

Zhou, Xueguang. 2000. "Reply: Beyond the Debate and Toward Substantive Institutional Analysis." *American Journal of Sociology* 105(4):1190–94.