Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

ELSEVIER Simulation Modelling Practice and Theory 10 (2002) 169–194

SIMULATION MODELLING PRACTICE AND THEORY

www.elsevier.com/locate/simpat

A pragmatic strategy for discovering and testing threats to the validity of sociotechnical experiments

William N. Dunn *

Graduate School of Public and International Affairs, University of Pittsburgh, 3R27 Posvar Hall, 15260 Pittsburgh, PA, USA

Received 23 May 2002

Abstract

This paper presents a new method for structuring decision problems as an essential aspect of solving them. This new method, the method of context validation, is a form of what Campbell (From Evolutionary Epistemology Via Selection Theory to a Sociology of Scientific Validity, 1996) and Dunn (Testing Rival Hypotheses with Pragmatic Eliminative Induction: The Case of National Maximum Speed Limits, unpublished; Am. Behav. Sci. 40 (3) (1997) 277; Knowledge, Power, and Participation in Environmental Policy Analysis, 2001) call pragmatic eliminative induction. By adding the term "pragmatic" to "eliminative induction," the method is distinguished from Mill's (J. Washington Acad. Sci. 16 (2) (1926) 317) influential analytical (or logical) variant of eliminative induction, which has been influential in designing social experiments that seek to investigate rival hypotheses, or so-called "threats to validity," that must be tested and eliminated to know whether a technological intervention is responsible for changes in a target social system. The paper shows how context validation can be used to estimate an approximately complete set of rival hypotheses, a *sine qua non* of research on complex sociotechnical systems. The method is exemplified by applying it to a major sociotechnical experiment, the US National Maximum Speed Limit of 1974. The method is shown to mitigate the commission of Type III errors (solving the wrong problem), which in the present case stemmed from the failure to define the relation between speed and traffic safety as a social and political, as well as technical problem. Because of this failure, policy makers concluded that maximum speed limits were effective in saving lives when they were not.

* Tel.: +1-412-648-7661; fax: +1-412-648-2605.
E-mail address: dunn@pitt.edu (W.N. Dunn).

1569-190X/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S1569-190X(02)00098-9

Keywords: Social experimentation; Quasi-experimental design; Sociotechnical systems; Problem structuring; Threats to validity; Pragmatic eliminative induction

1. Introduction

Every year thousands of technological innovations are designed with the aim of improving some social system. These innovations range from lasers and silicon chips to robots, communications networks, and systems for disposing of high-level nuclear waste. Although they may appear as purely technological, these and other innovations represent particular kinds of social experiments—sociotechnical experiments to be precise—because they involve the deliberate use of technology to change a social system. [1]

Sociotechnical experiments are almost always conducted in natural social systems. Rarely are they carried out in the comparatively closed, artificial systems of laboratories. Because they occur in natural systems, sociotechnical experiments involve many unmanageable contingencies that are beyond the control of designers. The ubiquity of these contingencies makes many sociotechnical experiments ''messy'' or ''ill-structured.'' In attempting to solve an ill-structured problem, the experimenter runs the risk of overlooking one or more of these contingencies, thereby omitting causally relevant factors that may be critical to the success of a sociotechnical intervention. The problem is not so much that ''nature'' is complex; it is that this very complexity is largely the product of conflicting beliefs about it.

Consider the sociotechnical problem of global warming. Some stakeholders—that is, individuals and groups with a stake in the problem because they affect or are affected by it—believe that global warming is an ''economic'' problem, and advocate potential economic solutions such as the auctioning of global pollution rights or the imposition of taxes and other disincentives. But other stakeholders define global warming as a ''political'' problem that calls for potential political solutions such as organized consumer boycotts or environmental insurrection. Still others define global warming as ''legal,'' ''cultural,'' ''technological,'' ''geophysical,'' ''biological,'' or ''ethical,'' advocating potential solutions which derive from their conflicting beliefs.

Under these conditions, standard models in decision theory, econometrics, and mathematical simulation may be of little value, because an adequate definition of the problem must be constructed before it is amenable to standard methods. To

[1] Portions of this paper draw on previous papers including [28,30–32]. My deep appreciation to Donald T. Campbell for reviews of early papers, and much else. In this paper my focus on ''sociotechnical experimentation'' is an effort to bring the methodology of social systems experimentation closer to work in engineering and the natural sciences. ''Sociotechnical design'' refers to a plan or strategy for assessing causal relations between technology and social systems by actively manipulating the initial conditions of an intervention (e.g., the choice of target and comparison groups) and selecting optimally appropriate methods of measurement, observation, and statistical analysis. The aim is always to maximize our capacity to detect causation, if it exists.

be sure, standard methods are appropriate and useful for solving relatively well-structured (deterministic) problems, where stakeholders, alternatives, outcomes, objectives, and values are known with certainty. Standard methods also may be appropriate for moderately structured problems (decisions under risk or uncertainty), where stakeholders, alternatives, outcomes, objectives, and values are known probabilistically. But standard methods are inappropriate and even useless in structuring ill-structured problems (decisions under ignorance). Here, the experimenter is like herbert Simon's architect who has been asked to design a custom house for which there are no standard plans [58]. Structuring the "right" problem is itself the problem. [2]

The purpose of this paper is to present and justify a new method for structuring problems as an essential aspect of solving them. I call this new method the method of context validation. I begin with a justification of context validation, which is a form of what Campbell [17] and Dunn [30–32] call pragmatic eliminative induction. By adding the term "pragmatic" to "eliminative induction," I distinguish it from John Stuart Mill's [44] classical analytical (or logical) variant of eliminative induction, which has had such a large impact on the design of experiments in agriculture, education, medicine, and other practical areas of application.

My focus is on the discovery of rival hypotheses—that is, alternative explanations of experimental outcomes—which must be tested and eliminated if we want to learn whether an intervention is responsible for changes in a target social system. The paper presents requirements for achieving approximate context validity, showing how the method of context validation can be used to estimate whether a collection of rival hypotheses is approximately complete. The method is exemplified by using it to assess the effectiveness of a major sociotechnical experiment, the US National Maximum Speed Limit of 1974.

My claim is that the method of context validation, had it been used by scientists and policy makers, could have averted the commission of a Type III error (solving the wrong problem), an error which stemmed from defining the relation between speed and traffic safety as a quintessentially technical problem. Policy makers could have avoided the commission of a Type I error (false positive), an error made by claiming that maximum speed limits is Europe and the United States were effective in saving lives when they were not.

2. Pragmatic eliminative induction

Unmanageable contingencies are rival hypotheses or, in the language of experimental and quasi-experimental design, threats to the validity of causal inferences

[2] Solving the wrong problem is a Type III error, as discussed Mitroff [45]. Type III errors are different from Type I and Type II errors, which involve "false negatives" and "false positives" that stem from setting the significance level too high or too low in testing the null hypothesis.

[22,24,56]. Threats to validity challenge claims that a sociotechnical intervention (the presumed cause) affects one or more outcomes (the presumed effects) by identifying and testing rival explanations.

2.1. Threats to validity

Traditionally, threats to validity were classified into two types: internal validity and external validity [22]. Subsequently, Cook and Campbell [24] provided two additional types, statistical conclusion validity and construct validity (see also Shadish et al. [56]). I have added another, called context validity. These five types of validity threats (rival hypotheses) are:

- *Threats to statistical conclusion validity*: The approximate validity of inferences about the covariation, in any of its statistical forms, between an intervention and one or more of its presumed outcomes. The approximate statistical conclusion validity of claims based on the classical linear model of regression analysis will be diminished to the extent that assumptions of linearity, homoscedasticity, uncorrelated errors, and other statistical requirements are violated.
- *Threats to internal validity*: The approximate validity of inferences about the existence of a causal relation between an intervention (the presumed cause) and one or more outcomes (the presumed effects), however statistically valid. The approximate internal validity of an inference relating cause and effect will be diminished to the extent that statistical covariation is weak or absent, the temporal precedence of the presumed cause is ambiguous or unknown, and other plausible causes are not eliminated.
- *Threats to external validity*: The approximate validity of inferences about the generalizability of internally valid causal relations to other contexts, settings, persons, groups, interventions, and outcomes. The approximate external validity of a generalized causal inference will be diminished to the extent that the effects of an intervention in one context or setting are undetectable in other contexts or settings, the original intervention is sufficiently complex (or diffuse) that its replication elsewhere is in doubt, and the outcomes are weak or absent among other persons or groups.
- *Threats to construct validity*: The validity of inferences about abstract categories, concepts, or labels used to characterize properties of contexts, settings, persons, groups, interventions, or outcomes, and one or more of their relations. The approximate construct validity of such categories, concepts, or labels will be diminished for reasons that include inadequate formal and operational definitions of constructs, failure to examine relations among multiple overlapping constructs, and failure to recognize and account for the effects of procedures for measuring and observing constructs on the existence of the constructs.
- *Threats to context validity*: The validity of inferences about the representativeness of causally relevant constructs, and hypotheses formed by these constructs, in specific social, spatial, and temporal contexts. The approximate context validity of constructs and hypotheses will be diminished to the extent that they are unrepre-

sentative of the conceptual ecology of persons who affect or are affected by an intervention. [3]

2.2. Classical eliminative induction

Threats to validity are tested and, where possible, ruled out through a process of *eliminative* induction. Eliminative induction has the prototypical form: "Initial observations of outcome y and intervention x confirm that y is (probably) the effect of x. However, additional observations of outcome y in the presence of contingency z confirm that y is (probably) the effect of z. Therefore, intervention x can (probably) be eliminated as the cause of y." By contrast, the prototype of *enumerative* induction is: "Repeated observations of outcome y and intervention x confirm that y is (probably) the effect of x".

The methodology of social experimentation is founded on eliminative induction [8,17]. [4] The methodology draws on Mill's methods for analyzing causes [44], especially his methods of agreement, difference, and concomitant variation. [5] The methodology nevertheless rejects any "essentialist" claim that it represents a way to achieve certainty in the discovery and validation of causes. And, although the methodology also builds on Karl Popper's theory of falsification in science [46,47], it recognizes that it is rarely if ever possible to falsify hypotheses conclusively. For this reason, the methodology is a qualified (limited) form of falsificationism which permits a critical examination of causally relevant contingencies, or rival hypotheses, that cannot be eliminated through randomization and other features of experimental design. [6] The number of these contingencies appears to be unmanageably huge, and the process of identifying and testing rival hypotheses is never complete [17].

How can we identify and test an unmanageably huge number of possible rival hypotheses? One potential answer is creativity and imagination. But creativity and imagination are difficult or impossible to teach, because there are no specific methods that ensure that creative acts can be repeated. Another solution is an appeal to

[3] The term "typical conceptual ecology" is similar to Egon Brunswik's "ecological validity," which was part of his theory of representative design (see [4]). Brunswik was referring more to what he and Edward Tolman called the "causal texture" of external environments and the role of perception in making judgments about them. For a recent essay on this problem see Dunn [33].

[4] Eliminative induction is not based on deductive reasoning, but on the process of inductively testing plausible rival hypotheses. It therefore differs from standard logico-deductive strategies designed to avoid infinite regress in the search for rival hypotheses. This strategy "solves" the problem of infinite regress by constructing a deductively valid argument that proceeds from a restricted hypothesis such as {X or Z or W cause Y}. If {X and Z} and {X and W} are observed as antecedents of Y, then {X causes Y}. This conclusion, based on Mill's method of agreement, is a logically valid argument form (*modus ponens*). See, e.g., Copi [25, pp. 371–377].

[5] Mill's methods are the methods of agreement, difference, agreement and difference, concomitant variation, and residues.

[6] Ostensibly, eliminative induction is redundant because the effects of rival factors (hypotheses) are eliminated by means of random selection of subjects, random assignment to experimental and control groups, and random assignment of the treatment to one of the groups.

"well-established" theories. However, the bulk of theories in the social sciences are disputed. "Well-established" theories are typically "well-defended" theories, and rival hypotheses are rarely entertained. [7] In this respect, reports on the practices of natural scientists are instructive [15, p. 353]. While natural scientists acknowledge that the range of rival hypotheses is potentially infinite, they pay little or no attention to this mere logical possibility. Rival hypotheses that might threaten the validity of some existing theory or hypothesis are rarely considered until logically possible rivals are transformed into empirically plausible ones and made available to the scientific community.

2.3. Pragmatism and the economy of research

This defensive posture toward rival hypotheses, memorialized by Imre Lakatos [36] with his well-known metaphor of a "protective belt," can be justified on grounds provided by an *economy of cognition* (*research*) principle: Choose a method that offers the most efficient way to achieve immediate and future research objectives. [8]

In the relatively simple world of the natural sciences, the economy of research principle can be used to justify a protective strategy towards rival hypotheses, a strategy that Lakatos [36] called "sophisticated falsificationism" to distinguish it from the "naïve falsificationism" he attributed to Popper [46]. But the principle is equally or more important in the more complex world of the social and behavioral sciences. The importance of the economy of research principle does not stem from theoretical or empirical considerations alone, but mainly from pragmatic concerns about the effectiveness and costs of conducting research in complex settings. Eliminative induction is a process of *pragmatic* (not logical or analytic) eliminative induction, and it is justified on grounds that it is the most efficient way to achieve research objectives in complex natural systems.

What I am calling pragmatic eliminative induction has a probative (erotetic) form as well as the protective one described above. In the protective form, the possibly unlimited character of rival hypotheses is acknowledged but ignored. Only firmly established rival hypotheses are judged suitable for testing. In this respect, it is a form of "sophisticated falsificationism," where the achievement of research objectives depends on success in protecting core beliefs. This requires long-term resistance to rival hypotheses which threaten the continued existence of a research program. The pragmatic justification for this kind of scientific protectionism may be illustrated with a familiar analogy: We search for the lost key under the street light because searching under the light is more likely to succeed than groping in the darkness.

"The darkness is greatest under the lamp," says a Hindi proverb. This counter-analogy calls attention to the need for a probative (erotetic) form of pragmatic eliminative induction. In this case, we attempt to *estimate* the completeness of rival

[7] This practice, which is similar to Lakatos's "sophisticated falsificationism" [38], may be elevated to the level of an erstwhile "rational choice."

[8] Charles Sanders Peirce [48] used the term "economy of research," rather than "economy of cognition." Both are similar to Zipf's "principle of least effort."

hypotheses in specific social, spatial, and temporal contexts. The process of estimation uncovers statistical regularities in the creation, distribution, and utilization of knowledge which have been identified and explained by Zipf's law of least effort [68], Bradford's law of scattering [1], Lotka's inverse square law of [50], and Merton's law of cumulative advantage in science [43]. All assert or imply that the distribution of information and knowledge conforms to a family of hyperbolic functions of the form [59]

$$x^n \cdot y = k$$

where y is the number of knowledge sources (e.g., authors), each making n contributions (e.g., scientific papers). Simon [57] has estimated parameters of these functions for the size distribution of the twenty largest world cities, the size distribution of industrial firms, and the distribution of rare and commonly used words in James Joyce's *Ulysses*, an example originally used by Zipf [68].

2.4. The social embodiment of knowledge

Pragmatic eliminative induction differs from the theories of Mill and Popper. These theories, as Rescher [52] argues, appear to require a metaphysical supposition that rival hypotheses reflect the "uniformity of nature" (Mill), or a notion that they originate in some super-sensory "convenient range structure." By contrast, pragmatic eliminative induction is based on a naturalistic epistemology which affirms that the only rival hypotheses at our disposal originate in *actually functioning* knowledge systems. All hypotheses—rival and friendly—are created, maintained, and changed within these natural systems. Hence, they are socially embodied, not just "out there," waiting to be discovered in a nature that is independent of observers. (Where else could they be embodied?)

Pragmatic eliminative induction strives to be comprehensive, without attempting the hopeless task of identifying and testing *all* rival hypotheses. As already noted, the probative (erotetic) form of pragmatic eliminative induction rejects strategies that confine the process of testing and elimination to a conveniently small number rival hypotheses. Although it is true that inductions are never complete, this fact alone is not sufficient to restrict the process of searching and testing by protecting already confirmed hypotheses. Among other difficulties, the protective strategy pre-empts opportunities for discovering omitted variables. Although even competent methodologists such as King et al. [36] seem to believe that there are precise econometric or statistical solutions to the "omitted variable" problem, the process of deciding which variables to include in a model is an imprecise and mixed enterprise involving the deduction of some hypotheses from existing theory, the abduction (retroduction) of others through imagination and creativity, and the induction of still others by "immersing" oneself in data. Although tests for specification error are readily available, the process of specifying the "right" model—especially, one that mitigates omitted variable bias—is essentially pragmatic. The process of conducting specification searches is limited by considerations of economy of cognition, and formal statistical tests play no role in that process. Statistical testing is fruitless when we are

unsure which variables should appear in a model, and, as Pindyk and Rubenfeld caution [49, pp. 162–164], we usually do not know the omitted variable.

3. The approximate completeness of rival hypotheses

Although our knowledge of rival hypotheses is never complete or certain, it is possible to *estimate* the limit of this range. We can seek rival hypotheses in the naturally occurring disagreements that take place within and between communities of scientists, policy makers, and citizens. Argumentation and public debate are an important source of rival hypotheses within these disputatious communities [12,26,35].

3.1. The problem of infinite regress

At first glance, the range of possible rival hypotheses seems unmanageably huge or infinite. Efforts to identify these rival hypotheses may seem like a never-ending process, a limitless search for ever more causally relevant factors. The problem is analogous to that of Morris Kline's [37] metaphorical homesteader clearing a plot of land in the wilderness. The homesteader is aware that enemies lurk in the wilderness that lies beyond the clearing. To increase his security, the homesteader clears a larger and larger area; but he is never fully confident that he is succeeding. Frequently, he must decide whether to clear more land, or trust the perimeter, where he must attained to the crops and livestock. The homesteader tries his best to push back the wilderness, knowing that the enemies that hide beyond the clearing may one day surprise and destroy him.

The point is that rival hypotheses cannot be ignored without consequences. To do so can and often does mean that the effects of causally relevant factors other than an intervention remain unknown, so that ineffective technologies are maintained. Another possibility is that interventions may appear harmful when in fact they are beneficial. Although restrictive eliminative induction does permit the identification and testing of some rival hypotheses—thus avoiding infinite regress in the search for rival hypotheses—this process is ad hoc, unsystematic, and biased. Here, the search for rival hypotheses is similar to the standard literature review, another ad hoc process that restricts the examination of studies to those which are readily available and consistent with the core beliefs of researchers. Indeed, the typical literature review is described by Light and Pillemer [39] as arbitrary, unsystematic, biased, and unscientific.

Pragmatic eliminative induction addresses this problem by building on the doctrine of fallibilism, as originally developed by the American pragmatist, Charles Sanders Peirce [48], and later by Karl Popper [46,47]. Fallibilism claims that much of the scientific knowledge we provisionally trust will eventually be found false, so that plausible scientific beliefs are those which have provisionally survived a process of elimination. Pragmatic eliminative induction nevertheless departs from elimination by falsification (falsificationism), because the process of elimination is not seen

as a direct confrontation of a dominant single theory with fact, but as a process of testing many competing theoretical and methodological hypotheses against available trusted "facts". [9] This pragmatic point of view rejects positivist principles of verification and correspondence, recognizing that theories are never unequivocally confirmed or falsified. It therefore rejects the notion that falsification demands the total overthrow of theories by pitting them against accumulated facts. The elimination of rival hypotheses is a matter of comparing the plausibility of one hypothesis against another, a process which Campbell observes, approaches but never fully attains completion [17, Sections 5.2–5.3].

3.2. Elimination through formal classification: the Campbell–Stanley–Cook–Shadish typology

The usefulness of the typology of threats to validity already presented above (Section 1) is that it increases the likelihood that rival hypotheses typically present in sociotechnical experiments will be incorporated in our analyses. Nevertheless, no procedures are available to estimate the extent to which the three dozen or more threats to validity detailed by Cook and Campbell [22] and Shadish et al. [56] provide a reasonable approximation of the real but unknown range of plausible rival hypotheses in given social, spatial, and temporal contexts. The typology is not sufficient for discovering this range, because to be sufficient, cautions Rescher [52], would require either a commitment to some a priori principle such as that of finite possibilities or truth-tropism, or substantive knowledge of a complete or approximately complete collection of rival hypotheses. In most cases, the typology assumes substantive knowledge of specific rival hypotheses which can be matched with the formal types and subtypes of threats to validity. This process of pattern matching must occur before the typology can be productively used. [10]

4. Requirements of context validation

I now want to elaborate on the fifth class of threats to validity already presented as a complement to the Campbell–Stanley–Cook–Shadish typology (Section 1). This fifth class is *threats to context validity*. It refers to the approximate validity of inferences about the representativeness of causally relevant constructs, and hypotheses formed by these constructs, in specific social, spatial, and temporal contexts. The method of context validation is guided by pragmatic criteria for making plausible

[9] It is as difficult to falsify as to corroborate a rival theory or hypothesis. Like many others, Campbell [7] and Cook and Campbell [24] invoke the Quine–Duhem thesis of the multiple theory-ladeness of observations to argue that single theories and single observations do not and cannot in practice conclusively falsify hypotheses. Popper held a similar qualified view of falsification.

[10] This point is similar to well-known critiques of Mill which have been available in standard logic texts for many years, e.g. [25].

truth-estimates through a process which Rescher [50] calls "cognitive systematization." Context validation has several requirements which include character, coordination, correctness-in-the-limit, and cost-effectiveness. [11]

4.1. Character

An estimate should have the same character as that which it estimates. Just as an estimate of a length should be a length, not a weight or temperature, an estimate of the proximal range of rival hypotheses should meet a character requirement: Rival hypotheses should be subjectively meaningful to, and elicited from, those who affect or are affected by sociotechnical interventions. The recognition that all knowledge is "embodied" leads to a search for rival hypotheses in naturalistic epistemologies, not in "nature" itself. For this reason, empirical estimates will violate the character requirement if they are not explicitly related to the concepts and hypotheses of those who make them. (It hardly needs saying that estimates are not *wholly* dependent on the subjective states or social circumstances of those doing the estimating.) Hence, estimates of context validity should focus on causally relevant concepts and hypotheses, not on their empirical referents. Estimates of the other types of validity—statistical conclusion, internal, external, construct—should and do focus on these referents. Concepts and hypotheses may be obtained by means of interviews, questionnaires, participant observations, ethnographies, and content analyses of texts which record subjectively meaningful conversations, arguments, debates, and discourses.

4.2. Coordination

An estimate of the proximal range of rival hypotheses should coordinate with the shape and distribution of beliefs in knowledge systems. The closer the coordination, the more accurate the estimate. An estimate of rival hypotheses within a disputatious community should be asymmetric and positively skewed, so that many persons share the same hypotheses and a few hold unique ones (see Fig. 1(a) and (b)). The *fewer* the plausible rival hypotheses, and hence the fewer the challenges to commonly held scientific beliefs, the greater the trust in the existing stock of knowledge. The *more* the plausible rival hypotheses, the greater the doubt about that knowledge. [12] Purposive

[11] These requirements draw on Rescher [52] and my later attempts [27,28,31] to adapt and incorporate them into a problem structuring methodology. Other relevant requirements include convenience, adaptability, uniformity, and generality.

[12] A high ratio of trust to doubt among leaders of scientific and political communities does not promote vigorous debate, creating a condition which Lindblom [40] calls "cognitive impairment." In this context, Campbell's conclusion [15, pp. 482–484] that there is a 99:1 trust-doubt ratio in most knowledge systems may actually understate the problem.

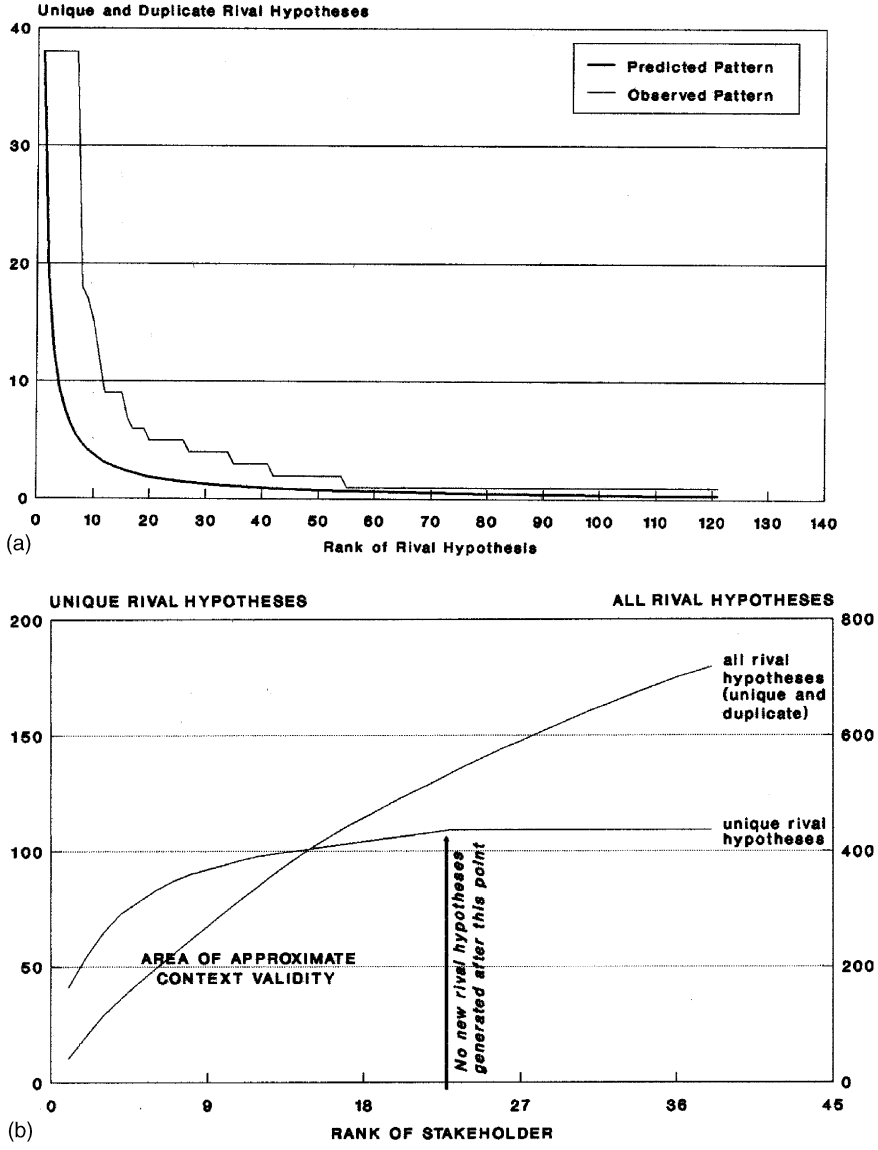

Fig. 1. (a) Observed pattern of approximate context validity predicted by Zipf's law of least effort, (b) area of approximate context validity predicted by Bradford's law of scattering.

sampling of persons who are linked socially and cognitively within the same knowledge system, rather than random sampling of unrelated individuals, is the appropriate procedure for selecting knowledgeables. A purposive sample is more likely to

produce the coordinated distribution of causally relevant concepts and rival hypotheses required for approximate context validity. [13]

4.3. Correctness-in-the-limit

As information on which an estimate is based becomes more complete, the estimate should eventually converge on the true range of rival hypotheses in a natural knowledge system. A cumulative frequency distribution of unique (non-duplicate) rival hypotheses may be arranged in order of decreasing frequency of occurrence and plotted on a graph. The plot of these unique rival hypotheses should eventually flatten out, indicating that we have achieved approximate context validity by approaching the limit of disputed truths in a natural knowledge system.

4.4. Cost-effectiveness

An estimate of the proximal range of rival hypotheses should be economical, in the sense of the "economy of cognition" principle already discussed. The total costs of obtaining an estimate will vary with the complexity of a knowledge system. [14] However, evidence from areas as diverse as marketing, linguistics, bibliometrics, operations research, statistics, and information science suggests by analogy that the range of a system of rival hypotheses can be approximated with observations obtained from a relatively small number of probes. [15] In experiments with small groups of knowledgeables in classroom and field settings, the latter including traffic safety, job training, and mine safety and health [32] the range of rival hypotheses has been reached in less than thirty probes. The probative value of a rival hypothesis—defined as its potential usefulness as a rival explanation—is inversely related to its frequency of occurrence. [16] The marginal costs of identifying each additional rival hypothesis increase at a diminishing rate, rapidly reaching a point where no new rival hypotheses are produced. The achievement of approximate context validity appears to be

[13] The metaphor of a laboratory meat grinder, attributed to Alan Barton, explains why random selection does not contribute to the requirement of coordination. Researchers investigating a particular strain of laboratory rat placed the animals in a meat grinder and subsequently analyzed a random sample of n elements of ground rat. They foolishly hoped that statistical analysis would yield not only unbiased statistical estimates (e.g., estimates of mean weight), but knowledge of rat anatomy and physiology, structural and functional features which cannot be (re)constructed through random sampling.

[14] Duncan MacRae, Jr. (personal communication) has shown that growth and decay curves in scientific citations are sensitive to the complexity of the field of research.

[15] Given a Bradford–Zipf–Mandelbrot distribution, the construction of a cost-effectiveness curve, displayed as an ogive, shows that the point of tangency (optimal efficiency) lies at about the 85th percentile. In the limit, an additional $0.15n$ probes, with n equal to the total number of stakeholders in a sample, would be necessary to obtain an approximately complete set of rival hypotheses.

[16] The theoretical justification for this claim originates in the application of information theory to theoretical linguistics. See Lyons [42].

cost-effective, reflecting an economy of cognition at the individual as well as knowledge-system level. [17]

5. The case of the national maximum speed limit

The importance of context validity to sociotechnical experimentation is evident when we probe the efficacy of the National Maximum Speed Limit of 1974, which established a uniform speed limit of 55 mph on all interstate highways in the United States.

5.1. Adopting the 55 mph speed limit

The 55 mph speed limit, adopted as a means to reduce gasoline consumption during the 1973–1974 OPEC oil embargo, was unexpectedly followed by a sharp decline in traffic fatalities. Between January 1 and December 31, 1974, there was a decline of 9100 fatalities, a 32% drop. Despite continuing opposition from rural Western states and the trucking industry, there was broad support for the policy among policy makers, policy analysts, and the general public. And, clearly, the problem is not unimportant or trivial. Traffic fatalities represent the leading cause of death among persons 35 years of age and younger, and average annual highway deaths since 1974 are equivalent to a fully loaded 767 aircraft crashing with no survivors every third day of the week.

5.2. Abandoning the 55 mph speed limit

On April 2, 1987, Congress enacted the Surface Transportation and Uniform Relocation Assistance Act of 1987, overriding President Reagan's veto. Provisions of this bill permitted individual states to experiment with speed limits up to 65 mph on rural interstate highways. By July, 1988 forty states had raised the speed limit to 60 or 65 mph on 89% of rural interstate roads. Senator John C. Danforth, an influential supporter of the 55 mph speed limit, argued against the new policy. The 65 mph speed limit, said Danforth, would save an average of one minute per day per driver, but result in an annual increase of 600 to 1000 deaths. [18] *The Washington Post* joined the opponents. "The equation is in minutes versus lives. It is not even close.... A hundred miles at 55 mph take about 17 minutes longer than at 65. That's the price of those lives. It ought to be the easiest vote the House takes this year." Eight years later, in a November, 1995 press release, US Secretary of Transportation Federico Pena reaffirmed the Clinton Administration's opposition to the higher speed limits. The National Maximum Speed Limit was officially abandoned the same year.

[17] Much like the "tragedy of the commons," economies at the individual level do not translate into economies at the collective level. Zipf's "principle of least effort" [68] may be applied at both levels, with very different consequences.

[18] Danforth's op-ed article appeared in The Atlanta Constitution (February 12, 1987).

The majority of analysts who have evaluated the effects of the intervention, along with elected officials from the ten northeastern states which until its repeal retained the 55 mph speed limit, have affirmed the conclusion that the 1974 law was responsible for the decline in traffic fatalities. However, the evidence shows that they failed to consider rival hypotheses which, had they been identified and tested, would have resulted in an altogether different explanation of traffic deaths and, consequently, a different policy recommendation. In effect, research on the 55 mph speed limit lacked approximate context validity.

5.3. Approximate context validity

To assess the approximate context validity of claims about the efficacy of the intervention, I conducted an analysis of documents prepared by 38 state officials responsible for documenting the effects of the 55 and 65 mph speed limits in their states. There is considerable political, administrative, professional, and geographic variation among these stakeholders, who include governors, secretaries of transportation, chief highway engineers, traffic safety analysts, and commanders of state highway patrols in every region of the country. [19] Because the sample is diverse and well-informed, it was anticipated that a large number of rival hypotheses would be identified.

As expected, there were sharp disagreements among the 38 stakeholders. Representatives of some states (e.g., Pennsylvania and New Jersey) were firmly committed to the hypothesis that the reduction of speeds to the maximum of 55 mph was causally related to the decline in fatalities. Others (e.g., Illinois, Washington, Idaho) were just as firmly opposed. Overall, 718 plausible [20] rival hypotheses were used by 38 stakeholders to affirm or dispute the effectiveness of the 55 mph speed limit in saving lives. Of this total, 109 hypotheses are unique because they do not duplicate hypotheses of any other stakeholder.

Table 1 lists the 109 unique rival hypotheses used to support or oppose the claim that the 55 mph speed limit was the principal cause of the decline in fatalities. The number of times a unique hypothesis was mentioned is shown in parentheses. Here, it is important to note that, from the standpoint of communications theory and language, the information-content of a hypothesis tends to be negatively related to its relative frequency, or probability of occurrence. [21] Hypotheses which are mentioned

[19] The twelve states not responding to the Department of Transportation request to review the DOT report include three northeastern states which retained the 55 mph speed limit (Connecticut, Maryland, Rhode Island). Other non-responding states changed to 65 mph on rural interstate highways in 1987 or 1988 (Arkansas, Indiana, Kansas, New Hampshire, New Mexico, Tennessee, Vermont, and Virginia).

[20] All rival hypotheses were judged to be plausible (i.e., they are taken to have a prior probability greater than zero). When they are generated in the manner described here, all rival hypotheses are plausible.

[21] The declining probative value of hypotheses (and of scientific trust) is explained by Lyons [42, p. 89] "Information-content varies inversely with probability. The more predictable a unit is, the less meaning it has. This principle is in accord with the commonly expressed view of writers on style, that cliches (or 'hackneyed expressions' and 'dead metaphors') are less effective than more 'original' turns of phrase."

Table 1
Rival hypotheses used by 38 state policy makers to oppose or support claims about the effects of the 55 mph speed limit in saving lives

The observed change (±) in fatalities after 1974 is caused by		
55 mph speed limit (38)	Speed adaptation (2)	Commuters (1)
Dual speed limits (38)	Interchange spacing (2)	Interchanges (1)
Law enforcement (38)	Stopping distance (2)	Travelers (1)
Average speeds (38)	Reaction time (2)	Driver alertness (1)
Dispersion of speeds (38)	Citation bias (2)	Overly easy driving
Driver education (38)	Age of drivers (2)	Environment (1)
Public information (38)	Weaving (2)	Awareness programs (1)
Unreliable data (18)	Driver alertness (2)	Injury rate (1)
Miscoded data (17)	Driver awareness (2)	Speed motivation (1)
Traffic volume (15)	Recovery zones (1)	Night driving (1)
Alcohol (12)	Lane restrictions (1)	Vehicle design (1)
Catastrophic accidents (11)	Perceived enforcement (1)	Property accidents (1)
Public support (10)	Number of lanes (1)	Risk taking (1)
Random fluctuations (9)	Length of mainlines (1)	Roadway safety (1)
Public attitudes (9)	Night speed limits (1)	Road conditions (1)
Sample size (8)	Equipment defects (1)	Energy crises (1)
"Safety switch" (7)	Acceleration lanes (1)	Non-linear relation of risk
State contexts (6)	Type of highway (1)	and speed (1)
Fatal accident rate (6)	Drinking age (1)	Terrain (1)
Trucks (5)	Commercial traffic (1)	Economic factors (1)
Accident histories (5)	Driver experience (1)	Safety diversion (1)
Interstate speed spillover (5)	Low-flying aircraft (1)	Contiguous roads (1)
Improper use fatality rates (5)	Truck accident rate (1)	Driver endurance (1)
Time of accidents (5)	Perceived fairness (1)	Population density (1)
Drugs (5)	Primary cause of crash (1)	Regression toward the
Highway design (4)	Police presence (1)	mean (1)
Speeding (4)	Tactical enforcement (1)	Police tolerance (1)
Weather (4)	Insurance premiums (1)	Augmented compliance(1)
Exceeding design speed (4)	Vehicle occupancy (1)	Vehicle proximity (1)
DUI cases (4)	Speed adaptation (1)	Passing maneuvers (1)
Driver error (4)	Economic mobility (1)	Collision accidents (1)
Traffic density (4)	Coefficient of variation (1)	Highway expansion (1)
Sanctions for violators (4)	Types of crashes (1)	Platooning (1)
Energy conservation (4)	Misuse of data (1)	Accident types (1)
Fluctuations in trends (3)	Enforcement contacts (1)	Accident exposure (1)
State autonomy (3)	Patrol hours (1)	Energy savings (1)
Vehicle safety (3)	Acceptable risk (1)	
Driver fatigue (3)	Availability of emergency	
Complexity of quantitative	Medical services (1)	
Variables (3)	Social costs (1)	
Construction projects (2)	Integrated compliance (1)	
Demonstration effect of	Declining oil prices (1)	
speeds (2)	Traffic controls (1)	
Failure to measure	Technological breakthroughs (1)	109 unique hypotheses
injury accidents (2)	child restraints (1)	718 total hypotheses

more frequently—those on which there is greater agreement or consensus—have *less* probative value than rarely mentioned hypotheses, because highly probable or

predictable hypotheses confirm what we already believe. They provide us with little or no information which might be useful in identifying potential rival explanations.

The rival hypotheses listed above were entered into a program designed for ethnographic research [22] and analyzed in two ways. First, all 718 rival hypotheses mentioned by the 38 stakeholders—unique rival hypotheses held by only one stakeholder as well as duplicates held by two or more stakeholders—were placed on a rank-frequency graph (Fig. 1(a)). This graph shows the entire pool of 718 hypotheses, which are ranked in descending order beginning with hypotheses with the highest frequency of occurrence. A second graph (Fig. 1(b)) shows the cumulative frequency distribution of unique (non-duplicate) rival hypotheses used by all stakeholders, beginning with the stakeholder generating the most rival hypotheses. There are 109 unique rival hypotheses.

As expected, the distribution in Fig. 1(a) conforms to the contours of Zipf's (1949) law of least effort, also known as his rank-frequency law. In turn, Fig. 1(b) conforms to the cumulative frequency distribution predicted by Bradford's law of scattering [1]. Similar frequency distributions have been found to characterize the structure of languages, cultures, information, and knowledge in a wide array of diverse natural systems. [23] In this respect, evidence accumulated over the past 200 years suggests that the structure of beliefs, ideas, and languages in natural systems of many kinds conforms to the theories of Lotka [41], Bradford [1], Zipf [68], and Price [50]. The extent of this conformity may be assessed by means of quantitative goodness-of-fit procedures, for example, linear regression analysis with a semi-logarithmic transformation (see, for example, Simon [57]). This conformity can also be evaluated visually, by matching the pattern of observations predicted by the theories to the actual pattern of observations [4,33]. [24]

The rank-frequency distribution displayed in Fig. 1(a) is similar to the patterns predicted by Lotka's inverse-square law of scientific productivity, Price's law of cumulative advantage in science, and Zipf's law of least effort in cognitive enterprises. In turn, the cumulative frequency distribution displayed in Fig. 1(b) is similar to Bradford's law of scattering, which has been used to estimate the redundancy of information in library holdings. The "empirical laws" of Lotka and Price may

[22] This shareware program is ANTHROPAC 3.2, developed by Stephen P. Borgatti, Department of Sociology, University of South Carolina, Columbia, SC, 29208.

[23] For a stimulating essay on these distributions, see Simon [57], who focuses on the work of Harvard socio-linguist George K. Zipf. The classic syntheses are provided in Zipf's studies of psycho-biology and language [66,67] and in his major work, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* [68]. Simultaneous and apparently independent work on variants of the same distribution include Lotka [41] and Bradford [1]. A thorough if somewhat dated review of work based on the Lotka–Bradford–Zipf distribution, along with a mathematical correction proposed by Benoit Mandelbrot, is Subramanyam [59]. In quantitative social studies of science, Derek de Solla Price [50] refines some of these distributions.

[24] On alternative procedures of pattern matching, see Dunn [33]. Regression analysis may be used to estimate the probability of a Type III error (testing the wrong rival hypotheses) by means of a coefficient of indetermination (equal to $1 - R^2$). Although regression estimates may be more accurate than visual ones, Tufte [61–63] argues persuasively that the two are complementary.

be explained by Zipf's "theoretical law" of least effort. The law of least effort also explains the principle of economy of cognition used as a basis for theories of induction developed by methodological pragmatists including Charles Sanders Pierce [48] and Nicholas Rescher [52,53]. Zipf's law states that human problem solving is based on the minimization of the total work required to solve immediate and future problems, as perceived by the problem solver. [25]

6. Threats to context validity

Any claim about the approximate context validity of a natural system of rival hypotheses is fallible, contingent, and corrigible. Accordingly, several threats to context validity may be invoked to challenge claims that we have obtained a valid estimate of the proximal range of rival hypotheses in a given social, spatial, or temporal context. Three threats to context validity are particularly relevant.

6.1. Inadequate responsiveness of methods

Methods used to elicit causally relevant constructs, and hypotheses formed by these constructs, may be more or less responsive to the character requirement. Inadequate responsiveness occurs when methods yield values or preferences, rather than causally relevant constructs, or when methods yield incorrect inferences about such constructs based on overt behavior or documents. Inadequate responsiveness of methods is a problem throughout the social and behavioral sciences. [26] In the present case, the open-coding of documents expressly prepared to address issues of causality appears to be responsive to the character requirement.

6.2. Discordance

The coordination of the observed distribution of rival hypotheses with the patterns predicted by the Lotka, Bradford, Price, and Zipf distributions may be more or less imperfect. If the sample had been restricted to a more homogeneous collection of stakeholders—for example, leading authors of journal articles on the 55 mph speed limit or like-minded advocates or opponents—the cumulative frequency distribution (Fig. 1(b)) would have flattened out prematurely. In the extreme case of total consensus on a single hypotheses, the cumulative frequency curve would show zero change after the first probe. The other extreme would be total disagreement, and the

[25] I have rephrased Zipf's law, which he states [68, p. 1] as follows. A person "will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems... as estimated by himself" (emphasis original). The Bradford distribution can be explained by Merton's "Mathew Effect" [43, pp. 221–278] and Price's principle of cumulative advantage in science [50, pp. 257–264].

[26] The determination of the "value" of information from its reported use, and the notion of "revealed preferences," are examples of the problem of inadequate responsiveness.

cumulative frequency curve would regress infinitely. In the present case, discordance does not appear to be a threat to context validity, given the proximal similarity of predicted and observed patterns. Although an expanded sample of stakeholders from the automobile industry, gasoline retailers, truckers, insurance companies, emergency medical services, and environmental protection groups might have provided additional rival hypotheses, the large number and variety of rival hypotheses elicited from these stakeholders suggests that there would be few changes in the present distribution.

6.3. Sub-optimal elevation

The elevation of the system of rival hypotheses, measured by the average number of constructs per stakeholder, may be suboptimal. Although precise judgments about optimal elevation would require prior knowledge of the complexity of the context of the intervention, it bears notice that it is the discovery of this complexity that is at issue. In the present case the elevation (mean) is approximately 15, given that 38 stakeholders mentioned a total of 718 hypotheses, of which 109 were unique. This number is far greater than that found in published literature on the 55 and 65 mph speed limits, a literature that displays little disagreement about the hypothesis that the 55 mph speed limit caused most of the observed decline in traffic fatalities after 1974. Sub-optimal elevation and discordance, which would be the result of excessive consensus and a premature flattening of the cumulative frequency curve, do not appear to be significant threats to context validity.

The approximately complete set of contextually valid rival hypotheses, already displayed in Fig. 1(a) and (b), contains many possible threats to internal, external, construct, and statistical conclusion validity. The most important of these are identified and tested below.

7. Threats to internal validity

Internal validity refers to the approximate validity of inferences about the existence of a causal relation between an intervention (the presumed cause) and one or more outcomes (the presumed effects), however statistically valid. The analysis of traffic fatalities, miles of travel, and employment during periods of economic recession and recovery suggests that business cycles affect changes in miles of travel and fatalities (Figs. 2 and 3). Three distinct cycles occur before and after recessions, and two of those recessions (1974–1975 and 1980–1982) followed sharp increases in oil prices by the Organization of Petroleum Exporting Countries (OPEC).

Fig. 2 poses a serious challenge to the hypothesis that the 55 mph speed limit was principally responsible for the observed decline in fatalities in 1974. Parallel conclusions about the effects of economic factors have been reached in research by Wilde [63] and Wilde and Simonet [64] examining the effect of unemployment on fatalities per 100,000 population in Europe and Canada. The negative correlation between unemployment and fatalities ranges from −0.68 to −0.88 [64, Chapter 5].

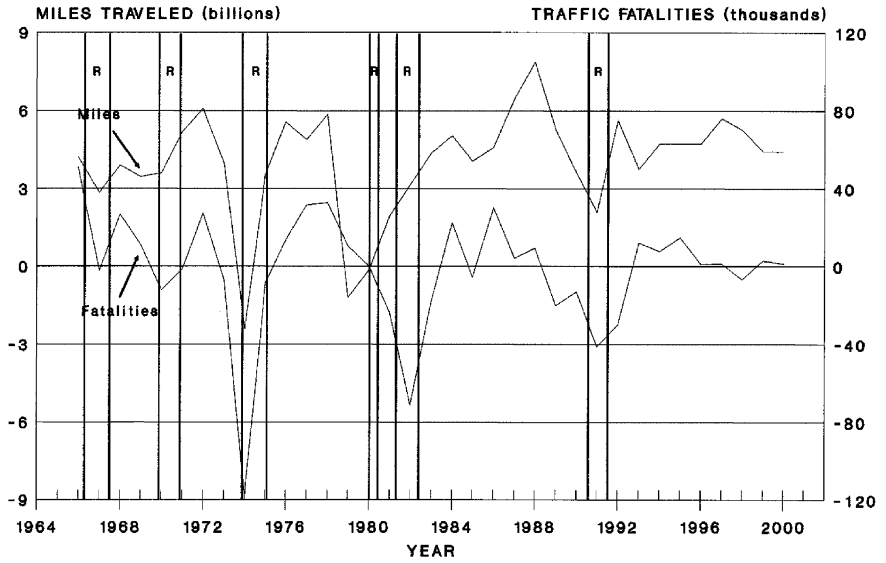

Fig. 2. Effects of recessions on annual changes in traffic fatalities and vehicle miles traveled, 1999–2000.

These conclusions accord with Fig. 3, which suggests that employment (rather than unemployment) positively affects vehicle miles traveled. [27] In turn, vehicle miles traveled affects fatalities.

Perhaps the most serious threat to internal validity is what may be called the "political economy hypothesis." Stakeholders who put forth this hypothesis contend that observed fluctuations in traffic fatalities over the 35 year period 1966–2001 can be explained by the state of the economies of the United States and European countries. These economies have been vulnerable to external petroleum shocks stemming from political conflicts between North Atlantic and Middle Eastern countries. This rival hypothesis, which falls in the formal category of *history* as a threat to validity, challenges the claim that the 55 mph speed limit caused the observed decline in fatalities by showing that one or more events occurring in the time interval between pre-policy and post-policy measurements affected the decline in fatalities. In this context, several stakeholders (see Table 1) argued that events including the 1973–1974 OPEC oil ambargo, energy crises, the price of petroleum, and other economic factors affected the decline in fatalities.

The line graphs showing relations among recessionary cycles, fatalities, miles driven, and employment (Figs. 2 and 3) supply a systemic, or molar causal, representation of the effects of the "exogenous economic factors," as these were described by one stakeholder. Additional analyses of relations among fatalities, vehicle miles traveled, and key economic indicators—including coincident, leading, and lagging economic indicators, industrial production, consumer confidence, and the price of

[27] The analysis of cross-correlations at lags of order 1–8 shows that the hypothesized effect of employment on miles driven is greater than the effect of miles driven on employment.

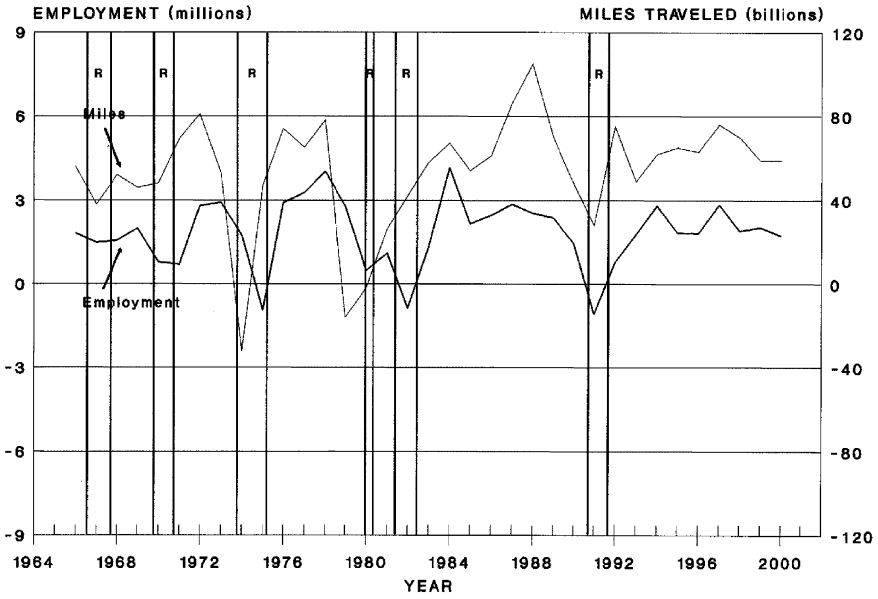

Fig. 3. Effects of annual changes in employment on annual changes in vehicle miles traveled, 1966–2000.

gasoline—yield simple correlation coefficients with the expected signs and time-series plots showing that each of variable tracks fatalities. The price of gasoline (in constant dollars) is negatively related to miles driven and fatalities, while coincidents economic indicators and consumer confidence have positive signs. As the state of the economy improves, vehicle miles driven and traffic fatalities increase. There is a strong positive cross-correlation between change in miles driven per cent (an index that combines fuel costs and fuel efficiency) and change in miles driven (Fig. 4). [28]

The hypothesis that the 55 mph speed limit affects fatalities may be challenged on still other grounds. [29] Stakeholders in 55 mph northeastern states identified the

[28] Miles per cent, a measure of average fuel cost, is based on *Social Indicators III* (December 1980), Table 4/23, p. 200. Miles per cent includes changes in the average cost of gasoline per gallon (constant 1982 dollars) and average fuel efficiency (EPA estimate).

[29] Space constraints do not permit a discussion of other threats to validity. For example, vehicle miles traveled are estimated on the basis of state excise taxes on gallons of gasoline sold, and different states keep more or less reliable records (instrumentation). Western states which abandoned the 55 mph limit have low population density and high fatality rates, while the reverse is true for northeastern states retaining the 55 mph limit. Population density, not speed limits, accounts for much of the difference (selection). Fatalities reached an extreme level (within the three regular cycles) immediately before the speed limit was increased in 1974 and just after it was abandoned in 1986–1987 (regression toward the mean). Part of the decline in fatalities is due to the "migration" of motorists from slower (and more dangerous) local roads to faster (and safer) interstate highways (mortality). Other rival hypotheses include mandatory seat belt laws and anti-drunk driving legislation (multiple treatment interference) and the growth of emergency medical services (history). Many of these rival hypotheses may account for some part of the long-term secular decline—but not for the observed cyclical variation.

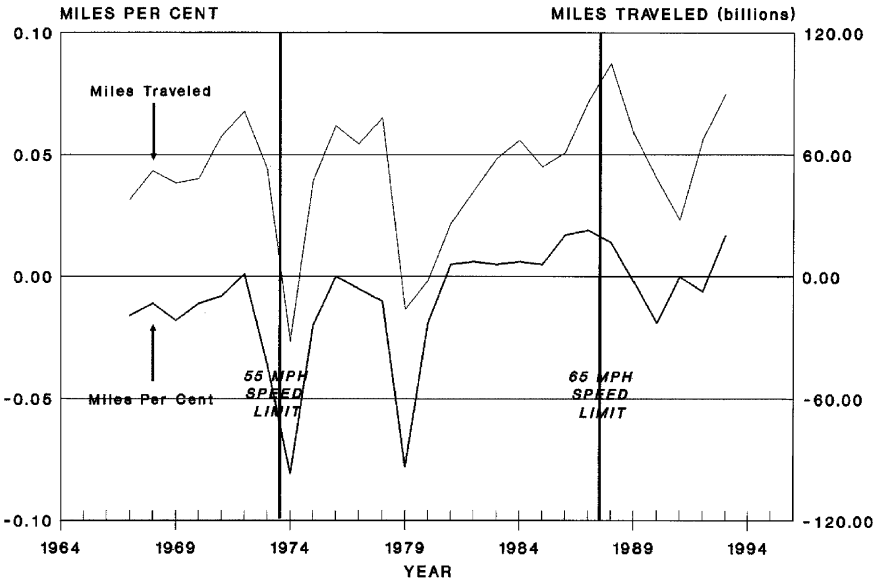

Fig. 4. Effects of annual changes in miles per cent on annual changes in vehicle miles traveled, 1966–1993.

"demonstration effect" on drivers in 55 mph states of changing the policy to 65 mph in the rest of the country. This threat to validity, the *imitation of treatments or policies* by controls, was eliminated because according to the National Highway Traffic Safety Administration, fatalities in 55 mph states were not responsive to the small increase in average speeds from 58.7 to 58.9 mph between 1986 and 1987. Stakeholders in the 55 mph states also identified rival hypotheses of "speed spillover" from 65 to 55 mph states and a perceived sense of "unfairness" among drivers in 55 mph states, which represent threats to validity formally known as *diffusion of treatments or policies*, *compensatory equalization of treatments or policies*, and *resentful demoralization of controls* who have been deprived of a perceived benefit. Given the relatively small increase in average speeds in 55 mph states, none of these rival hypotheses appears plausible. Indeed, these rival hypotheses may be somewhat beside the point, given cross-sectional analyses which report no statistically significant correlation between average speeds and fatality rates.

To estimate the joint effects of employment, miles of travel, and the 55 mph speed limit on fatalities a MARIMA (multivariate autoregresive integrated moving average) model was estimated. The model was useful in representing interactions among variables and in correcting for autocorrelation and random instability (white noise) in the time series. The most economical MARIMA model employs miles of travel, employment, and the 55 mph speed limit (a dummy variable) to explain traffic fatalities. The coefficient which measures the effect of the 55 mph speed limit is statistically significant ($p = 0.01$) only for the year immediately following the implementation of the 55 mph speed limit. The coefficients for miles of travel and employment are

statistically significant ($p = 0.05$) over the entire period in which the three recessionary cycles occur. The coefficient of determination is strong ($r^2 = 0.96$) and statistically significant ($p = 0.001$).

7.1. Threats to construct validity

Theoretical constructs may be inadequately conceptualized, defined, and measured, creating the impression of an adequate theoretical explanation when none exists. Concepts such as speed or velocity, while they are essential in theories and theoretical laws in physics, explain the severity of impact—*not* fatalities and fatal accidents. Commonsense understandings such as "speed kills" reinforce the invalidity of such explanations. Average speeds and the standard deviation of these speeds have increased since 1974, but the long-term secular trend in otherwise cyclically fluctuating fatalities is one of continuing decline.

7.2. Threats to statistical conclusion validity

In the case of the 55 mph speed limit, there are three potential threats to statistical conclusion validity. The instability characteristic of random fluctuations in the time series must be removed by logarithmic transformations of one or more variables and/or the inclusion of a moving average term. If instability is not removed, there is a bias toward Type II errors. When observations in the time series are not indepen-

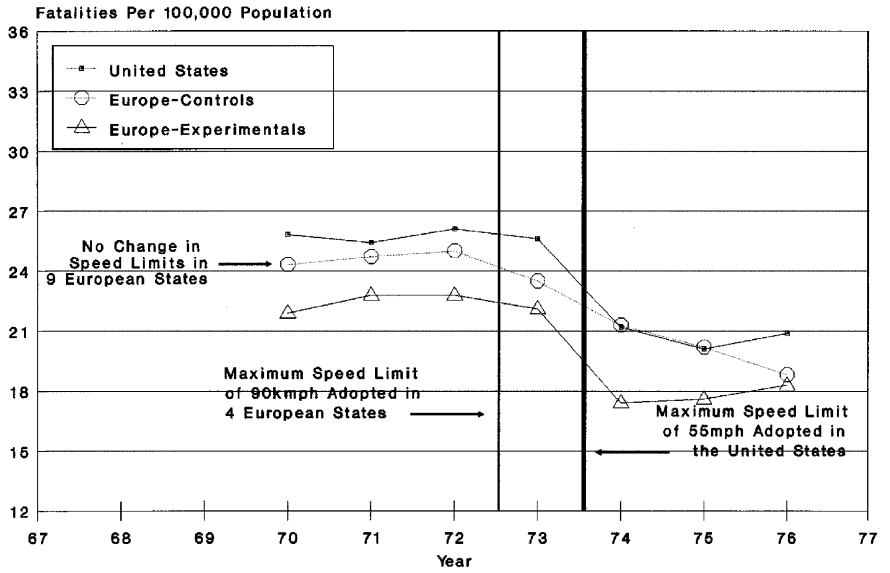

Fig. 5. Maximum speed limits have weak temporary effects on fatality rates in Europe and the United States.

dent, auto-correlation will inflate a significance test, suggesting that the 55 mph has an effect when it does not. The inclusion of an autoregression term in a MARIMA model permits appropriate significance testing and adequate statistical conclusion validity. A MARIMA model requires a minimum of 50–60 observations. Otherwise there is a risk of a Type I error. The three threats to statistical conclusion validity were eliminated by differencing the time series and conducting a sensitivity analysis of the effects of a sample size. Tests for autocorrelation and the effect of sample size showed that the results are statistically sound.

7.3. Threats to external validity

A control series analysis (Fig. 5) was performed to assess the approximate external validity of the political economy hypothesis. The United States was compared with European countries which imposed new maximum speed limits of 48–54 mph (80–90 kph) after 1973, as well as countries which did not impose new limits. A similar pattern of declining fatalities per 100,000 population occurred in the United States and in European countries with and without a new maximum speed limit. This suggests that the political economy hypothesis applies in Western Europe as well as the United States. The OPEC oil embargo and its effects on gasoline prices, and on recessionary cycles, provides a compelling explanation of traffic fatalities. [30] Although the 55 mph speed limit had a temporary effect which eroded after 1975, most of the variation in fatalities is explained by political and economic factors associated with the international politics of oil and domestic business cycles.

Space constraints do not permit a discussion of other threats to validity. But these are important. For example, vehicle miles traveled are estimated on the basis of state excise taxes on gallons of gasoline sold, and different states keep more or less reliable records (instrumentation). Western states which abandoned the 55 mph limit have low population density and high fatality rates, while the reverse is true for northeastern states retaining the 55 mph limit. Population density, not speed limits, accounts for much of the difference (selection). Fatalities reached an extreme level (within the three regular cycles) immediately before the speed limit was increased in 1974 and just after it was abandoned in 1986–1987 (regression toward the mean). Part of the decline in fatalities is due to the "migration" of motorists from slower (and more dangerous) local roads to faster (and safer) interstate highways (mortality). Other rival hypotheses include mandatory seat belt laws and anti-drunk driving legislation (multiple treatment interference) and the growth of emergency medical services (history). Many of these rival hypotheses may account for some part of the long-term secular decline—but not for the observed cyclical variation.

[30] Campbell (personal communication) noted that the results presented here are equally or more definitive than findings he and others have reported with respect to similar social experiments including [54,55].

8. Conclusion

This paper has attempted to show how a pragmatic strategy of eliminative induction satisfies four requirements of context validity: character, correctness-in-the-limit, coordination, and cost-effectiveness. The intuitively appealing hypothesis that speed limits save lives is rejected and replaced with what I have called the political economy hypothesis. The conclusion that domestic and international economic factors are principally responsible for cyclical fluctuations in traffic fatalities is unlikely to have been discovered and tested without the results achieved through context validation. Although pragmatic eliminative induction is always incomplete, historically dated, and fallible, it nevertheless permits us to identify an approximately complete set of rival hypotheses, a *sine qua non* of research on the effects of sociotechnical interventions [2,3,5,6,9–11,13,14,16,18–21,23,29,34,51,60,65].

References

[1] S.C. Bradford, Sources of information on specific subjects, Engineering 137 (1934) 85–86.
[2] D.T. Campbell, Factors relevant to the validity of experiments in social settings, Psychological Bulletin 54 (1957) 297–312.
[3] D.T. Campbell, Methodological suggestions from a comparative psychology of knowledge processes, Inquiry 2 (1959) 152–182.
[4] D.T. Campbell, Pattern matching as an essential in distal knowing, in: K.R. Hammond (Ed.), The Psychology of Egon Brunswik, Holt, Rinehart and Winston, New York, NY, 1966, pp. 81–106.
[5] D.T. Campbell, Definitional versus Multiple Operationalism, vol. 2, 1969, pp. 14–17, SP, 31–36.
[6] D.T. Campbell, Reforms as experiments, American Psychologist 24 (1969) 409–429, SP, 261–289.
[7] D.T. Campbell, Evolutionary epistemology, in: P.A. Schilpp (Ed.), The Philosophy of Karl Popper, Open Court Press, Lasalle, IL, 1974, pp. 413–463, SP, 393–434.
[8] D.T. Campbell, Quasi-Experimental Designs, 1974, in: SP, 191–221.
[9] D.T. Campbell, Degrees of freedom and the case study, Comparative Political Studies 8 (1975) 178–193, SP, 377–388.
[10] D.T. Campbell, Descriptive Epistemology: Psychological, Sociological, and Evolutionary. William James Lectures, Harvard University, 1977, SP, 435–486.
[11] D.T. Campbell, Qualitative knowing in action research, in: M. Brenner, P. Marsh, M. Brenner (Eds.), The Social Context of Method, Croom Helm, London, 1978, pp. 184–209, SP, 360–376.
[12] D.T. Campbell, Experiments as arguments, Knowledge: Creation, Diffusion, Utilization 3 (1982) 327–337.
[13] D.T. Campbell, Relabelling internal and external validity for applied social scientists, in: W.M.K. Trochim (Ed.), Advances in Quasi-Experimental Design and Analysis, Jossey-Bass, San Francisco, CA, 1986, pp. 67–77.
[14] D.T. Campbell, Guidelines for monitoring the scientific competence of preventive intervention research centers: an exercise in the sociology of scientific validity, in: National Policies for Optimizing Validity in Applied Social Science, in: W.N. Dunn (Ed.), Knowledge: Creation, Diffusion, Utilization, 8, no. 4, March 1987, pp. 389–430.
[15] D.T. Campbell, in: E.S. Overman (Ed.), Methodology and Epistemology for Social Science: Selected Papers, University of Chicago Press, Chicago, IL, 1988.
[16] D.T. Campbell, Systems Theory and Social Experimentation, Unpublished Paper prepared for the USSR–US Conference on Systems Theory and Management, May, 1988, US coordinator: Prof. S.A. Umpleby, Management Science, George Washington University.

[17] D.T. Campbell, From evolutionary epistemology via selection theory to a sociology of scientific validity, 1996. Revised and edited by Barbara Frankel and Cecilia Heyes. Also in Evolution and Cognition, 1997.
[18] D.T. Campbell, The experimenting society, in: W.N. Dunn (Ed.), The Experimenting Society: Essays in Honor of Donald T. Campbell, vol. 11 of Policy Studies Review Annual, Transaction Books, New Brunswick, NJ, 1998.
[19] D.T. Campbell, D.W. Fiske, Convergent and discriminant validation by the multitrait–multimethod matrix, Psychological Bulletin 56 (1959) 81–105, SP, 37–61.
[20] D.T. Campbell, R.F. Boruch, Making the case for randomized assignment to treatments by considering the alternatives: six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects, in: C.A. Bennett, A. Lumsdaine (Eds.), Evaluation and Experiment: Some Critical Issues in Assessing Social Programs, Academic Press, New York, 1975, pp. 195–296.
[21] D.T. Campbell, H.L. Ross, The Connecticut crackdown on speeding: time-series data in quasi-experimental analysis, Law and Society Review 3 (1968) 33–53, SP, 222–237.
[22] D.T. Campbell, J.C. Stanley, Experimental and Quasi-Experimental Designs for Research, Rand McNally, Chicago, IL, 1963.
[23] T.D. Cook, Postpositivist critical multiplism, in: R.L. Shotland, M.M. Mark (Eds.), Social Science and Social Policy, Sage Publications, Beverly Hills, CA, 1985, pp. 21–62.
[24] T.D. Cook, D.T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings, Houghton Mifflin, Boston, MA, 1979.
[25] I.M. Copi, Introduction to Logic, Macmillan, New York, 1953.
[26] W.N. Dunn, Reforms as arguments, Knowledge: Creation, Diffusion, Utilization 3 (1982) 293–326.
[27] W.N. Dunn, Methods of the second type: coping with the wilderness of conventional policy analysis, Policy Studies Review 7 (4) (1988) 720–737.
[28] W.N. Dunn, Discovering the boundaries of ignorance in policy research: a continuous measure of errors of the third type, in: B.M.J. Wolters, K.M. Schultz, J.T.A. Koster, F.L. Leeuw (Eds.), Between Sociology and Sociological Practice: Essays on Social Policy Research—Liber Amicorum Dedicated to Mark van de Vall, Institute for Applied Social Sciences, Nijmegen, 1993, pp. 36–62.
[29] W.N. Dunn, Policy reforms as arguments, in: F. Fischer, J. Forester (Eds.), The Argumentative Turn in Policy Analysis and Planning, Duke University Press, Durham, NC, 1993, pp. 254–290.
[30] W.N. Dunn, Testing rival hypotheses with pragmatic eliminative induction: the case of national maximum speed limits, Graduate School of Public and International Affairs, Pittsburgh, PA, unpublished paper, January 1995.
[31] W.N. Dunn, Probing the boundaries of ignorance in policy analysis, American Behavioral Scientist 40 (3) (1997) 277–298.
[32] W.N. Dunn, Using the method of context validation to mitigate type III errors in environmental policy analysis, in: M. Hisschemoeller, R. Hoppe, J.R. Ravetz, W.N. Dunn (Eds.), Knowledge, Power, and Participation in Environmental Policy Analysis, vol. 12 of Policy Studies Review Annual, 2001.
[33] W.N. Dunn, Pattern matching: methodology, in: T.D. Cook, C.C. Ragin (Eds.), The International Encyclopedia of the Social and Behavioral Sciences Part 2, Section 3, Logic of Inquiry and Research Design, Elsevier, New York, 2002.
[34] F. Fischer, J. Forester (Eds.), The Argumentative Turn in Policy Analysis and Planning, Duke University Press, Durham, NC, 1993.
[35] R. Hoppe, Political Judgment and the Policy Cycle: The Case of Ethnicity Policy Arguments in the Netherlands, in: F. Fischer, J. Forester (Eds.), The Argumentative Turn in Policy Analysis and Planning, Duke University Press, Durham, NC, 1993.
[36] G. King, R. Keohane, S. Verba, Designing Social Inquiry, Princeton University Press, Princeton, NJ, 1994.
[37] M. Kline, Mathematics: The Loss of Certainty, Oxford University Press, New York, NY, 1980.
[38] I. Lakatos, in: J. Worrall, G. Currie (Eds.), The Methodology of Scientific Research Programmes, Cambridge University Press, Cambridge, UK, 1978.

[39] R.J. Light, D.B. Pillemer, Summing Up: The Science of Reviewing Research, Harvard University Press, Cambridge, MA, 1984.
[40] C.E. Lindblom, Inquiry and Change: The Troubled Attempt to Understand and Shape Society, Yale University Press, New Haven, CN, 1990.
[41] A.J. Lotka, The frequency distribution of scientific productivity, Journal of the Washington Academy of Sciences 16 (2) (1926) 317–323.
[42] P. Lyons, Theoretical Linguistics, Oxford University Press, New York, 1969.
[43] R.K. Merton, in: N.W. Storer (Ed.), The Sociology of Science, University of Chicago Press, Chicago, IL, 1973.
[44] J.S. Mill, A System of Logic, Metheun, London, 1843.
[45] I.I. Mitroff, The Subjective Side of Science: A Philosophical Inquiry into the Psychology of the Apollo Moon Scientists, Elsevier, New York, NY, 1974.
[46] K.R. Popper, Conjectures and Refutations, Basic Books, New York, NY, 1963.
[47] K.R. Popper, The Logic of Scientific Discovery, Basic Books, New York, NY, 1968.
[48] C.S. Peirce, in: C. Hartshorne, P. Weiss (Eds.), Collected Papers of C.S. Peirce, Harvard University Press, Cambridge, MA, 1931.
[49] R.S. Pindyck, D.L. Rubinfeld, Economic Models and Econometric Forecasts, third ed., McGraw-Hill, New York, NY, 1991.
[50] D.S. Price, Little Science, Big Science... and Beyond, Columbia University Press, New York, 1986.
[51] W.V.O. Quine, Word and Object, MIT Press, Cambridge, MA, 1960.
[52] N. Rescher, Induction, University of Pittsburgh Press, Pittsburgh, PA, 1980.
[53] N. Rescher, The Limits of Science, University of California Press, Berkeley, CA, 1984.
[54] H.L. Ross, D.T. Campbell, The Connecticut speed crackdown: a study of the effects of legal change, in: H.L. Ross (Ed.), Perspectives on the Social Order: Readings in Sociology, McGraw-Hill, New York, 1968, pp. 33–53.
[55] H.L. Ross, D.T. Campbell, G.V. Glass, Determining the social effects of a legal reform: the British Breathalyser crackdown of 1967, American Behavioral Scientist 13 (1970) 493–509.
[56] W. Shadish, T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton Mifflin, Boston, MA, 2002.
[57] H.A. Simon, The sizes of things, in: J. Tanur, et al. (Eds.), Statistics: A Guide to the Unknown, Holden Day, New York, 1972.
[58] H.A. Simon, The structure of ill structured problems, Artificial Intelligence 4 (1973) 181–201.
[59] K. Subramanyam, Laws of scattering, in: A. Kent (Ed.), Encyclopedia of Information Science, Marcel Dekker, New York, 1975, pp. 336–354.
[60] E.R. Tufte, Data Analysis for Politics and Policy, Prentice Hall, Englewood Cliffs, NJ, 1974.
[61] E.R. Tufte, The Visual Display of Quantitative Information, Graphics Press, Cheshire, CN, 1983.
[62] E.R. Tufte, Visual and Statistical Thinking: Displays of Evidence for Making Decisions, Graphics Press, Cheshire, CT, 1997.
[63] E.R. Tufte, Visual Explanations: Images and Quantities, Evidence and Nature, Graphics Press, Cheshire, CN, 1997.
[64] G.J.S. Wilde, Target Risk: Dealing with the Danger of Death, Disease and Damage in Everyday Decisions, PDE Publications, Toronto, 1994.
[65] G.J.S. Wilde, S.L. Simonet, Economic Fluctuations and the Traffic Accident Rate in Switzerland: A Longitudinal Perspective, Swiss Council for Accident Prevention, Berne, 1996.
[66] G.K. Zipf, Selected Studies in the Principle of Relative Frequency of Language, Harvard University Press, Cambridge, MA, 1932.
[67] G.K. Zipf, Psycho-Biology of Language, Houghton Mifflin, Boston, MA, 1935.
[68] G.K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Chicago, IL, 1949.