Introduction

I.1 The idea of a controlled social experiment

I.1.1 *The historical origin of the income maintenance experiments*

Beginning in the early 1960s, the work of quantitative social scientists and applied statisticians began to play an increasing role in social policy deliberations in the United States. But their experience with the Coleman Report, which appeared in mid-1966, made many leading policy-oriented statisticians and social scientists doubt the value of large-scale observational studies (such as Coleman's) as aids in formulating social policy. Some began to argue for the statistician's classical Fisher-type controlled experiment as the needed precursor to public policy formulation.[1]

Meanwhile, rebellions in the inner cities replaced the civil rights movement of the early 1960s. In an economic expansion driven by the Vietnam War, these rebellions triggered one traditional response of the state: an expansion of the welfare system and a stepping-up of the "War on Poverty." But political liberals in the antipoverty programs and welfare rights groups began to argue for even more: a guaranteed income program to replace the demeaning welfare system. Economists of many political persuasions were amenable, for different reasons, to one form of guaranteed income: a negative income tax (NIT). But political conservatives in Congress balked, claiming that such a program would lead the poor to stop working.

In late 1966, Heather Ross, an M.I.T. economics doctoral candidate, working for a Washington antipoverty agency, made a proposal that eventually broke the political deadlock (Ross 1966). Aware of statisticians' reservations about observational studies, Ross suggested a controlled social experiment to test congressional conservatives' claim that

[1] Gilbert and Mosteller (1972) presented an influential argument for "controlled experiments."

an NIT would lead the poor to stop working. The proposal gained the assent of both liberals and conservatives and eventually grew into the four income maintenance experiments of 1968–80.

Researchers planned the four studies as random-assignment experiments like those routinely employed in agricultural and biomedical empirical inquiry. In agronomy, treatment-group fields receive specified quantities of fertilizer, and researchers compare treatment- and control-group crop yields. In biomedicine, treatment-group patients receive specified drug doses, and researchers compare treatment- and control-group responses. So in the income maintenance experiments, treatment-group families would receive an NIT with a (guarantee level, tax rate) amount, and researchers would compare the hours worked for the treatment and control groups. (Later they decided to compare divorce rates too.) Though the income maintenance researchers usually thought of themselves as methodological pioneers, there had been precedents for the use of random-assignment experiments in the study of social phenomena. For example, in the 1950s, psychologists studied "juvenile delinquents" in Cambridge-Somerville, Massachusetts. The treatment group received psychological counseling, and researchers compared treatment- and control-group recidivism rates.[2]

As data collection and analysis proceeded, much of the income maintenance research gradually moved away from its planners' original conception. For example, agricultural and biomedical experiments usually compare the mean responses of several randomization-created treatment groups. Instead, the income maintenance researchers began to employ linear (or even nonlinear) response models. Sometimes they modeled two responses simultaneously. And with the hours-worked response variable, the researchers often related the model's functional form to a utility function and budget constraint of the microeconomic theory of labor–leisure choice. Ultimately, the connection of much of the income maintenance data analysis to the randomization-created groups became obscure.[3]

[2] For a reevaluation of this study, see McCord and McCord (1959).

[3] For movement to a response model approach (i.e., "parameterized treatment") in the experiments' planning stage, see Orcutt and Orcutt (1968). Hausman and Wise (1979) provide an example of a simultaneous-equations or structural model. Keeley et al. (1978) provide an example of work based on the theory of labor–leisure choice. The work of Robins and West (1980) contains a simultaneous-equations or structural model and is also based on the theory of labor–leisure choice.

I.1.2 *Two senses of "experiment"*

Statisticians usually claim that a random-assignment study is a more reliable basis for inferring cause and effect than is an observational study. For example, Gilbert and Mosteller argue:

> . . . suppose that we take a large group of people, give a randomly chosen half of them $1000 a year and give the other half nothing. If at the end of two years we find that the half with the extra money have increased the size of their dwellings considerably more than the others, few will argue that the increased income was not a cause. The random assignment of the experimental treatment triumphs over all the little excuses for not attributing the effect to the difference in treatment. (1972, p. 372)

Yet here and elsewhere statisticians usually leave vague what makes a random-assignment study a more reliable basis for inferring cause and effect than is an observational study.

In fact, the analysis that statisticians conduct with data from random-assignment studies often fails to involve any assumptions which depend on randomization. For example, in a simple (i.e., unstratified) random-assignment study, statisticians usually assume that two independent normal probability distributions with the same variances generated the treatment and control observations. Then they conduct analysis of variance – for example, test the hypothesis that the two normal-distribution means are equal. Yet such analysis nowhere depends on the fact that randomization occurred: One would draw the same inferential conclusions from the analysis had random assignment not occurred, that is, had the study been an observational one.

Part I of this book seeks to clarify the relationship between random assignment and cause-and-effect conclusions. First we construct an inferential logic based on the observations and random assignment. Then we uncover difficulties for such a logic posed by social phenomena in particular. Finally we construct a second inferential logic based on the observations and infinite populations, but not on randomization. Thus, we seek at once to construct a randomization-based logic, lay bare its conceptual difficulties, and classify the difficulties into those which inhere in the logic and those which appear only in its application to social phenomena. Hence, Part I is an immanent critique of social experiments in the random-assignment sense of "experiment."

However, "experiment" also can mean an empirical study which gets a set of observations to confront and possibly falsify a theory. Because some of the income maintenance research involved the microeconomic theory of labor–leisure choice, perhaps the four studies were "experiments" in this second, theory-testing sense. Further, if the theory-testing sense of "experiment" characterizes science more generally, then perhaps the income maintenance experiments were tests of a scientific theory of microeconomics.

In fact, however, many have in recent years questioned that the falsification of theories through tests even properly characterizes the natural sciences, let alone social sciences. If these skeptics are correct, then understanding the income maintenance studies as theory-testing experiments seems a hopeless task; that is, it is a way of misunderstanding them. Further, there is a body of microeconomic theory and empirical work – including the income maintenance studies – which *conceivably* conforms to a theory-testing logic. But there is another body of microeconomic work – including general equilibrium theory and stability analysis – which seems to follow an entirely different logic.

Part II seeks to construct two logics of economic inquiry. First we catalogue difficulties which arise in reconstructing both natural science and economics as generation and testing-falsification of theories. Presumably, none of these difficulties keeps economics from being science, because they don't keep natural science from being science. Then we search for difficulties which arise in reconstructing economics, but not natural science, as generation and testing-falsification of theories. *Conceivably* these latter difficulties include some which keep economics from being science. Finally, we study how the logic of general equilibrium theory and stability analysis differs from, and is related to, a theory-testing logic. Hence, Part II is an immanent critique of controlled social experiments as tests of theory.

I.1.3 *The role of the income maintenance experiments in this book*

The income maintenance experiments originated in the social conflict and urban unrest and upheavals of the 1960s and early 1970s. Dissipation of the unrest and transformation of the conflict followed the experiments' completion. So one could seek to understand the experiments in their historical context. Such a historical inquiry might reasonably ask: What social groups favored conducting the experiments and why? or, What distortions arose in transmission of the experimenters' findings to the political realm? or, Did the reported findings cause the eventual non-adoption of an NIT? And understanding the experiments' relation to subsequent events would require detailed knowledge of precise findings and results.

We do *not* seek such a historical understanding of the income maintenance experiments here.[4] Instead, our goal is a critical conceptual and methodological understanding of two modes of social inquiry. We focus on the income maintenance experiments as examples of those two modes of social inquiry. For our methodological inquiry, we ask, What sorts of hypotheses can randomization aid in testing, and what concept of cause and effect is implicit in such tests? or, How do controlled social experiments resemble and differ from agronomy and physics experiments? or, Are there unavoidable difficulties for controlled social experiments which make their results unreliable? We are less interested in the details of the income maintenance experiments' results and findings than in the details of their techniques as examples of the two modes' methods of social inquiry.

We begin the investigation of each mode of *social* inquiry by asking what are the most *general* conceptual difficulties of the mode's methods; that is, we ask, How is *any* randomization-based inference conceptually problematic? or, How does *any* science fail to be theory-testing experiments? Next we ask, Are there any *special* difficulties for a mode of inquiry applied in the *social* realm? Here we distinguish unavoidable difficulties in applying a mode's methods to social phenomena from experimenter errors. For example, the random-assignment mode's Hawthorne effect is unavoidable. But stratified assignment without a weighting of the strata in analyzing the data is an experimenter error. Or, the theory-testing mode's ad hoc models of differences between individuals are unavoidable given the present form of microeconomics. But the use of a labor-supply function which presupposes utility *minimization* is an experimenter error.

Part II also explores this question: Is microeconomics science in the same sense that natural sciences such as physics are? We note that microeconomic experiments such as those on income maintenance seem to share a set of conceptual difficulties with natural science experiments such as those of classical mechanics. Hence, those shared difficulties can't be a basis for concluding that microeconomics is not science. But microeconomics experiments such as those on income maintenance also exhibit a set of conceptual difficulties not evident in natural science experi-

[4] Neuberg (1986; in press) begins a historical critique of the income maintenance experiments.

ments such as those of classical mechanics. So the second set of conceptual difficulties *may* be a basis for concluding that microeconomics in its present form is an undeveloped science.

I.2 The book's structure

I.2.1 *The argument by chapters*

Part I develops a critical account of the statistical logics of controlled social experiments. Chapter 1 examines the philosophical basis of a first statistical logic of controlled experiments. We develop two distinct concepts of cause and effect and two of probability, and then we fuse one of the cause-and-effect concepts and one of the probability concepts into a notion of a gambling-device law (GDL). Chapter 2 demonstrates that the GDL notion is the philosophical basis of a first statistical logic of controlled experiments by constructing a theory of randomization-based causal inference based on it. The theory of causal inference includes an unbiased point estimate of the mean causal effect, a test of the hypothesis of no causal effects, and a test of the hypothesis of no mean causal effect. We also extend the theory of unbiased causal point estimation to a case of univariate regression. And throughout the chapter we focus on the theory's conceptual difficulties and stress the role of assumptions (in addition to the assumption that the observations are randomization-generated) in the inference.

Chapter 3 introduces the technique of double-blindness to the theory of unbiased causal point estimation and so extends the theory to situations involving a conscious object of inquiry. But Chapter 3 also stresses that certain features of a controlled *social* experiment introduce unavoidable biases into the causal point estimation of Chapter 2. For example, participants are blind neither to being in an experiment nor to their treatment, and they are volunteers and free to drop out of the experiment before its completion. And these features lead to Hawthorne effect, self-selection, and attrition biases. Chapter 4 develops the second statistical logic of controlled social experimentation. This second logic – a Neyman–Pearson theory logic – is based on the concept of an infinite population. We stress that randomization plays no role in the second logic (e.g., the logic's inferential statements based on data from a controlled social experiment would be the same had an observational study generated the data). And we note that weakening the infinite-population assumption of the second logic leads to weaker inferential conclusions.

Part II develops a critical account of two microeconomic logics, one of which is sometimes involved in controlled social experiments. Chapter 5 constructs a theory-testing logic of classical mechanics experiments. Then we note some difficulties (e.g., circular-like reasoning, nonfalsifiable laws, and approximation conventions) with an account of classical mechanics experiments as theory-testing. Chapter 6 constructs a theory-testing logic of microeconomic experiments, like those on income maintenance. Then we note precisely the same difficulties with an account of microeconomics as theory-testing, as with Chapter 5's account of classical mechanics as theory-testing. But we also note a second set of difficulties, which have no classical mechanical parallels, with an account of microeconomics as theory-testing. This second set of difficulties includes how to know that one observes a market equilibrium and how to deal with differences in individual consumer tastes and production technologies. If microeconomics fails to be science, then, it is because of the second set of difficulties, not because of those philosophical problems it shares with natural sciences like classical mechanics.

Chapter 7 considers the income maintenance studies as theory-testing microeconomic experiments and discovers the second set of difficulties catalogued in Chapter 6. But we also note a third set of difficulties – avoidable researcher errors. These include a failure to employ a labor-supply function that follows deductively from maximization of a utility function, and a failure to use an estimating procedure with statistical optimality properties. Chapter 8 develops the second microeconomic logic – a logic of general equilibrium theory and stability analysis. We construct this logic as a normative argument for a perfectly competitive private-ownership economy. Among the logic's difficulties, we note two of special importance. First, the logic fails to prove that a perfectly competitive private-ownership economy leads in any given amount of time to a Pareto-efficient allocation which is in the interest of society. Second, the logic fails to discover an assumption set on consumer preferences consistent with both introspection and stability analysis in its present form.

The Conclusion reconsiders in further depth those conceptual difficulties of controlled social experiments which are neither researcher errors nor general philosophical difficulties of science. We classify these middle-range or methodological difficulties into three families: the problems of complexity, behavior, and time. The problem of behavior includes the difficulties of atomization, individuality, consciousness, and will. The problem of time includes the difficulties of dynamics and equilibrium, time-dependent utility-production, and history. Chomsky's linguistics is an example of a social science which avoids the problems of behavior and time and is mastering the problem of complexity. We conclude that controlled social experiments in their present form are undeveloped science and are somewhat overrated tools for public policy decision-making.

I.2.2 *The reasons for formalization*

In the course of the book's argument we consider five distinct logics in some detail. Chapters 2 and 3 develop a randomization-based logic and Chapter 4 an infinite-population-based logic of statistical inference for data from controlled experiments. Chapters 5 and 6 construct logics of theory-testing experiments in classical mechanics and microeconomics, respectively. Chapter 8 develops two versions of the logic of a microeconomic argument for a perfectly competitive private-ownership economy. Although it is the custom to do so only for the two statistical logics and the last microeconomic logic, we present each of the five logics as a formalism. That is, we develop each logic as a quasi-axiomatic system of axioms, definitions, theorems, and so forth. Why do we stress formalization more than is usual?

Many economists claim that their work bears some resemblance to that of physicists. We therefore formalize the theory-testing logics of classical mechanics and microeconomics so as to be able to most carefully appraise the claim that economics and physics are similar. With both theories expressed in quasi-axiomatic form, similarities appear more sharply. For example, formalization reveals that the central law sketches in each theory follow deductively from optimization procedures. That is, Newton's Second Law in classical mechanics follows from God's or nature's minimization of the action (possibly subject to constraints). And the marginal utility and product conditions in microeconomics follow from consumer–producer maximization of utility-production functions subject to constraints.

But with both theories expressed in quasi-axiomatic form, differences also appear most sharply. For example, formalization reveals that in classical mechanics there is a single law sketch which when coupled with a particular force equation implies a prediction statement which can be confronted with observations. In contrast, formalization reveals that in microeconomics, many law sketches, when coupled with particular utility and production function equations, imply particular market supply and demand equations. These market supply and demand equations, in turn, imply prediction statements which can be confronted with observations.

The mathematician Kurt Gödel proved that any axiomatization of ordinary arithmetic could not be both complete and consistent.[5] Gödel's proof suggests that any attempt to put knowledge on the formal basis of an axiom system may run into conceptual difficulties. Each of our five logics is simultaneously a mathematical system and a set of words meant to say something about the world of our observations. So we formalize the logics in part to uncover and catalogue the conceptual difficulties which Gödel's proof suggests we will encounter. That is, formalizing a logic serves to clarify what aspects of the logic resist formalization. Hence, formalization enables us to compare logics not only on the basis of their formal structures but also on the basis of which of their aspects resist formalization.

For example, formalizing randomization-based tests of the hypothesis of no causal effects reveals that such tests depend on a formalization-resistant choice of test statistic; or formalizing theory-testing logics of classical mechanics and microeconomics identifies formalization-resistant points and suggests tentative conclusions about barriers to microeconomics becoming fully developed science; or, finally, formalizing frequently made normative arguments for a perfectly competitive private-ownership economy allows us to identify gaps and inconsistencies in those arguments.

I.2.3 *Some recurring themes*

The book's overall theme is a critique of the methods of controlled social experimentation. On this theme we argue by demonstration that if the methods are laid out in detail, their weaknesses will emerge from *within* that account of the techniques. For example, we make no argument against randomization in controlled experiments as some others have done.[6] Rather, we construct a formal logic for randomization-based inference in controlled experiments, and from within that logic itself the fact emerges that testing the hypothesis of no mean causal effect is problem-

[5] See Gödel (1970) for Gödel's proof and Findley (1942) for a conceptual explication of the proof.

[6] The earliest direct argument against randomization was made by the statistician Gosset ("Student" 1937). More recently, some philosophers have revived Gosset's original argument (Seidenfeld 1981; Levi 1983).

atic. Hence, our account of the methods of controlled social experimentation is an *immanent* critique.

In Part I we develop the theme that, contrary to the beliefs or hopes of many social statistical researchers and statisticians, empirical statistical inference always depends on assumptions which go beyond the observations. For example, randomization-based inference depends on the metaphysical assumption of deterministic causality; or our second logic of statistical inference depends upon the assumption of the existence of unobservable infinite populations. We call the need for assumptions, in statistical inference, that go beyond the observations the problem of empiricism.

In Part II we develop the theme that, contrary to the beliefs of some philosophers, a theory-testing logic of science fails to solve such "epistemological problems" as the problem of empiricism. Rather, in the theory-testing logic, the problem of empiricism becomes the problem of how words in axioms and theorems hook onto the world of our observations. For example, attaching the word "mass" in classical mechanics or the words "supply" and "demand" in microeconomics to the world of our observations involves circular-like reasoning.

Closely related to the problem of empiricism is another theme which runs contrary to the understanding of many statisticians and economists: Weakening the assumptions of statistical inference or economic theory weakens the inferential or deductive conclusions one may draw. For example, Chapter 4 shows that weakening the classical normal-probability-distributions assumption in a test of the hypothesis that two population means are equal makes significance probabilities asymptotic in the inferential conclusion. That is, one can speak, under the hypothesis, not of how likely the observations were, but only about how likely they would have been in the limit as sample sizes approached infinity. Chapter 8 shows that weakening the assumption of consumer-preference strong convexity to convexity weakens the argument for a perfectly competitive private-ownership economy. That is, the argument becomes this: Such an economy is consistent with, rather than eventually leading to, a Pareto-efficient allocation.

In Part II and in the Conclusion we pursue the theme of parallels in classical mechanics and microeconomics. We don't *directly* deny the claim of some economists that economics is a science of society on a par with physics' account of nature. Rather, in immanently critical fashion we develop certain similarities between classical mechanics and microeconomics. Where the similarities leave off then demonstrates that to the