1

Introduction

1.1 Economics as an experimental discipline

> One possible way of figuring out economic laws . . . is by *controlled experiments*. . . . Economists [unfortunately] . . . cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe. (Samuelson and Nordhaus, 1985, p. 8)

Samuelson and Nordhaus echo a widely shared view that some disciplines are inherently experimental, but others (including economics) are not. History has not been kind to this view. In Aristotle's day some 2,000 years ago, even physics was considered nonexperimental. About 400 years ago, innovators such as Bacon and Galileo established a tradition of controlled experiments, mostly in physics. Experiments in related disciplines such as chemistry followed. For a long time biology was considered inherently nonexperimental because its subject was living organisms, but Mendel, Pasteur, and others introduced new experimental techniques in the nineteenth century. Modern biology certainly is an experimental science. Even psychology, whose mental subject matter might seem least accessible to laboratory study, has evolved a distinctive experimental tradition over the last century.

History suggests that a discipline becomes experimental *when innovators develop techniques for conducting relevant experiments.* The process can be contagious, with advances in experimental technique in one discipline inspiring advances elsewhere. Still, each discipline must innovate for itself. Even closely related disciplines differ in their intellectual focus, so wholesale transfer of experimental technique across disciplinary boundaries is seldom possible.

It took a long time but economics has finally become an experimental science. Most economists have heard about the experimental work of Vernon Smith, Charles Plott, Reinhard Selten, and others in the last three decades. (Indeed, in later editions of their text Nordhaus and Samuelson edited out the remarks we quoted.) Experiments are now commonplace in industrial organization, game theory, finance, public choice, and most other microeconomic fields. Some aspects of macroeconomic theory recently have been examined experimentally, although full-scale macroeconomic experiments do not seem feasible for budgetary and political reasons. (We refer to true, controlled experiments; uncontrolled macroeconomic "experiments" are all too common in recent years!) Perhaps macroeconomics too, like meteorology and astronomy, will become an indirectly experimental discipline, one that relies on experimentally verified results in constructing its central theories, although the central theories themselves are not amenable to direct experimental examination.

The methods as well as the substance of experimental economics are new in some respects. In the last few years the substantial findings of experimental economics have been expertly surveyed; see the annotated bibliography in Appendix I, pp. 143–74. However, no readily accessible, self-contained summary of experimental method and technique has yet been written for students and researchers in economics. The purpose of this primer is to bridge that gap.

Chapters 2 through 8 examine specific methods and techniques for economic experiments. The final chapter takes a look at the emergence of experimental economics in the last thirty years. The present chapter touches on some preliminary but fundamental issues: the interaction between theory and empirics, the differences between experimental and nonexperimental data for empirical work, and the diverse purposes of experiments. Since this book is a primer and not a theoretical treatise, we barely skim the surface of the deeper philosophical issues.

1.2 The engine of scientific progress

Theory organizes our knowledge and helps us predict behavior in new situations. In particular, theory tells us what data are worth gathering and suggests ways to analyze new data. As theory progresses, it guides us in refining our use of data and in selecting questions we should ask.

Conversely, data collection and analysis often turn up regularities that are not explained by existing theory. Such empirical regularities spur refinement of theory, usually as minor adjustments and sometimes as revolutionary changes. Kuhn (1970) and Lakatos (1978) discuss how

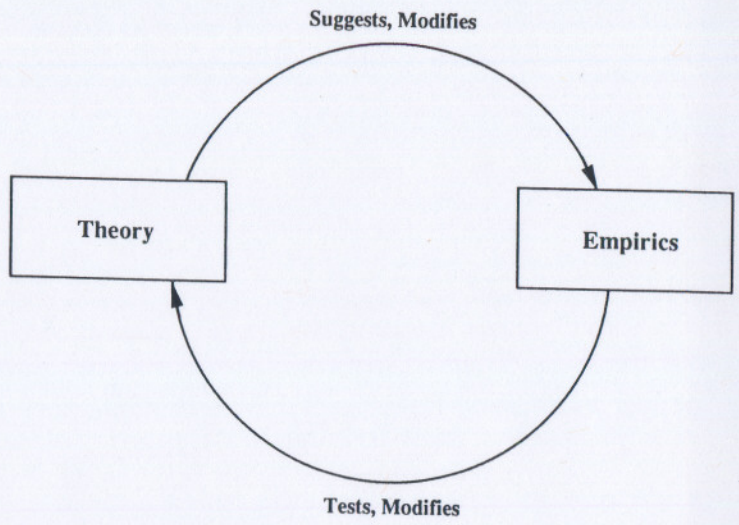

Fig. 1.1 Theory and empirics.

data and theory interact over time. The alternation of theory and empirical work, each refining the other, is the engine of progress in every scientific discipline. (See Figure 1.1.) Economics is no exception. Traditionally, observations from naturally occurring economic phenomena were the only source of data to stimulate revision of theory. If data relevant to an economic proposition could not be captured from naturally occurring conditions, then the proposition went without benefit of empirical refinement. In recent years, experimental methods have given economists access to new sources of data and have enlarged the set of economic propositions on which data can be brought to bear.

1.3 Data sources

Data for empirical work can be drawn from several types of sources, each with distinctive characteristics, advantages, and disadvantages. A key distinction is between *experimental data,* which are deliberately created for scientific (or other) purposes under *controlled* conditions, and *happenstance data,* which are a by-product of ongoing *uncontrolled* processes. A less important but still useful distinction can be drawn between *laboratory data,* which are gathered in an artificial environment designed for scientific (or other) purposes, and *field data,* which are gathered in a naturally occurring environment.

All combinations are possible. For example, an experimenter may

	Happenstance	Experimental
Field	Rate of Inflation in U.S.	Income Maintenance Experiments
Laboratory	Discovery of Penicillin	Laboratory Asset Markets

Fig. 1.2 Examples of data sources.

intervene in a naturally occurring process and record the outcomes; such data are field-experimental (FE). An economic example is the income-maintenance experiments in Denver, Seattle, and elsewhere (see Killingsworth, 1983; Pencavel, 1986). Traditionally, almost all empirical work in economics has used field-happenstance (FH) data such as national income accounts, commodity prices, or corporate financial statements. The story goes that penicillin was discovered in a laboratory when controls failed in a nutrient experiment, so this is an example of rare laboratory-happenstance (LH) data. Of course, this primer focuses on the last type of data, laboratory-experimental (LE). In this and later chapters, we often loosely refer to LE data as laboratory data or as experimental data and often ignore LH and FE data, but we make the finer distinctions when necessary.

Experimental data (LE or FE) are especially valuable for scientific purposes because they are relatively easy to interpret. If outcome Y (say, highly efficient allocations) is always associated with institution X (say, a certain kind of auction market) as institutional and other environmental variables are manipulated in a well-designed experiment, then we can confidently conclude that X causes Y. Happenstance data can't support such confident causal conclusions. Given the absence of control, an observed correlation between X and Y may be due to Y indirectly causing X, or may be due to some unobserved variable Z causing both X and Y. Leamer (1983, p. 31) makes the point while satirizing Monetarists and Keynesians in his delightful "Luminist versus Aviophile" parable. Aviophiles explain the higher crop yields found under trees in terms of bird droppings, while Luminists explain the same finding in terms of light intensity. Their quarrel is unresolvable with the "field" data because the two explanatory variables are completely confounded – that is, shade and bird droppings go together. The process-control example in Box, Hunter, and Hunter (1978, p. 487ff) provides a more elaborate discussion of the same point. We defer discussion of the underlying statistical issues until Chapter 7.

The other main issues in comparing experimental and happenstance data are cost and validity. Flexible, controllable laboratory environments usually are expensive to build, maintain, and operate, and each experiment requires further costs such as payments to human subjects. Thus both fixed (or sunk) costs and marginal costs may be significant for laboratory experiments, and typically are even higher for field experiments. Of course, it is also costly to obtain new field-happenstance data. The costs of gathering FH data on individual choice behavior, for example, are about the same as for LE data. Obviously it is least expensive to use data previously collected by someone else, such as a government agency.

Validity (or relevance) is a crucial issue for all data sources. When the field environment is of direct interest, FH and FE data are automatically relevant. On the other hand, FH data are normally,

> collected by government or private agencies for non-scientific purposes. . . . [By contrast,] astronomers are directly responsible for the scientific credibility of their data in a way that economists have not been. In economics, when things appear not to turn out as expected the quality of the [FH] data is more likely to be questioned. . . . (Smith, 1987, p. 242)

Specifically, the validity of FH data often is impaired by the omission of the really interesting variables (necessitating use of crude proxies), by measurement error of unknown magnitude, or by skewed coverage.

Laboratory data pose different validity questions. First, there is the question of internal validity: Do the data permit correct causal inferences? As we will see in later chapters, internal validity is a matter of proper experimental controls, experimental design, and data analysis. Second, there is the question of external validity: Can we generalize our inferences from laboratory to field? The issue of external validity or relevance often troubles economists who are unfamiliar with experimental work, and it remains a concern for experimentalists. Chapter 2 begins with a discussion of the gentle art of designing relevant experiments. Parallelism, the last substantive topic in Chapter 2, deals directly with the general question of external validity. For now, suffice it to say that, in economics as in other experimental disciplines, external validity has been firmly established in a diverse set of laboratory studies.

Sometimes data from computer simulations or surveys are improperly labeled as experimental economic data. Computer simulations of a theoretical model (no human decision makers involved except in writing the computer code) are best regarded as a type of theoretical results rather than as empirical data. Traditionally the investigator uses deductive logic

and mathematical derivations to discover the implications of a theoretical model. You may resort to simulation because you have an intractable theoretical model so you can't derive the relevant theorems. As computing power becomes cheaper and more convenient, computer simulations become increasingly attractive relative to formal derivations as a discovery method. Survey data (human responses to hypothetical questions) are empirical but, unless responses are economically motivated, their reliability as economic data is questionable. This last point is developed in Section 2.3.

1.3.1 Some evidence

Econometricians have devised many ingenious techniques to deal with the weaknesses of happenstance data. Direct opportunities to test the effectiveness of these techniques are rare, LaLonde (1986) being the prime example. (See Cox and Oaxaca, 1991, for a different kind of effectiveness test.) LaLonde obtained field-experimental earnings data on former participants and nonparticipants in a job-training program. Experimental control had been achieved by random assignment of individuals as participants or nonparticipants; this important technique is discussed in Section 3.2. Straightforward statistical procedures showed that participants' mean annual earnings were about $900 higher, a statistically significant difference.

LaLonde then treated the data as if it were happenstance and the "control group" of nonparticipants did not exist. He used standard data sources and several multiequation specifications (some involving self-selection) and several econometric procedures to estimate the earnings effect. Estimates of the job-training effect on earnings varied considerably and some even had the wrong sign. He concludes

> This study shows that many of the econometric procedures and comparison groups used to evaluate employment and training programs would not have yielded accurate or precise estimates of the impact of the National Supported Work Program. The econometric estimates often differ significantly from the experimental results. Moreover, even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results. (LaLonde, 1986, p. 617)

The point is that, when obtainable at comparable cost, experimental data allow more reliable inferences than happenstance data. There are many cases where happenstance data are adequate and cheap; then experiments are not worthwhile. In many other cases happenstance data are inadequate and experimental data can be obtained at reasonable cost. Such cases present the best opportunities for experimental work.

Different types of data can be complementary. You can combine evidence from computer simulations, field, and laboratory to get sharper conclusions than those obtainable from a single data source.

1.4 **Purposes of experiments**

Experiments have many possible purposes. The proper way to design and to conduct your experiment depends on your purpose. Before proceeding further, a review of the purposes of experiments is in order (see Plott, 1982, 1987).

Some experiments have been conducted to generate data that might influence a specific decision. For example, Grether and Plott (1984) report an experiment designed to provide evidence in an antitrust case. Hong and Plott's (1982) research arose from a case considered by Interstate Commerce Commission. Alger (1988), Alger, O'Neill, and Toman (1987a,b), Plott (1988), and Rassenti, Reynolds, and Smith (1988) discuss the experiments conducted to assist Federal Energy Regulatory Commission. Roth (1987a) refers to experimentation designed to influence policymakers as "whispering in the ears of princes."

Influencing authorities is not the only persuasive purpose for experiments. Innumerable laboratory and field experiments have been conducted in order to provide data on how best to influence the decisions of consumers, voters, and managers. Cohen (1992) reports that white American consumers are more responsive to advertisements for stereo equipment featuring Asian models. This responsiveness of demand for stereos, where Asian manufacturers have dominated the U.S. market, is not discernible in advertisements for pickup trucks. Recently several popular business magazines have discussed new field technology that allows accurate measurement of market response to product innovations or advertising campaigns. In U.S. presidential campaigns at least since 1988, laboratory studies of voter response to proposed television messages and campaign slogans have played an important part in the strategies of most major candidates. For example, Torry and Stencel (1992) report in the *Washington Post* that the Bush–Quayle campaign confirmed through focus groups that bashing trial lawyers was an effective vote-getting theme; see Payne (1992) for another typical example. The large (and apparently increasing) sums of money devoted to such marketing applications suggests that they do provide commercially valuable data.

This primer emphasizes the scientific purposes of experiments. Persuasion certainly is still in the picture (McCloskey, 1985), but specific immediate decisions are of less concern than the longer run views of

the scientific community. One scientific purpose is to discover empirical regularities in areas for which existing theory has little to say. McCabe, Rassenti, and Smith (1993) and Friedman (1993), for example, compare the properties of several market institutions whose theoretical properties are as yet poorly understood. Smith (1982b) calls such experiments *heuristic*. In other areas, by contrast, several competing theories offer differing predictions and experiments can help map the range of applicability for each theory. For example, Fiorina and Plott (1978) study committee decisions in the laboratory and find that only a few of the sixteen models and variants considered are at all consistent with the data. Finally, there are areas for which only one model is applicable. Laboratory work can demonstrate whether there are any conditions under which the theory can account for the data, and if so, can test theory for robustness. "In Search of Predatory Pricing," by Isaac and Smith (1985) is a negative example. Smith (1982b) refers to the last two types of experiments as *boundary* experiments and refers to sets of experiments intended to establish definitive broad laws of behavior as *nomothetic*.

Some experimental economists have hesitated in recent years to describe the purpose of an experiment as a *test of theory*. From a formal point of view, a theory consists of a set of axioms or assumptions and definitions, together with the conclusions that logically follow from them. A theory is formally valid if it is internally consistent – that is, it does not lead to statements that contradict each other – and if the conclusions are indeed provable from the assumptions. What can be learned about theories by conducting experiments? Some experimentalists (including most psychologists) think of experimental data as a means of testing the descriptive validity of the assumptions about human behavior on which the theory is based. Others (including most economists) would readily grant that the behavioral assumptions of most economic theories do not and need not meet the descriptive validity criterion used in psychology. Instead they believe that a theory is of direct practical interest only to the extent that its conclusions provide good approximations (relative to alternative theories) of actual behavior even when its assumptions are not precisely satisfied. See Friedman (1953) and Koopmans (1957) for further discussion.

The proper job of the empirical scientist is to find regularities in observed behavior in a broad range of interesting environments and to see which theories can best account for these regularities. Whether this job is called "testing theories," or more circumspectly referred to as "seeing which theories best organize the data," it is a primary purpose of scientific experiments.

Experimental economists have become increasingly interested in recent years in using laboratory methods (including economic incentives) to measure individual (innate or "home-grown") characteristics in the population, such as willingness to pay for environmental amenities or risk aversion (see Cummings, Harrison, and Rutström, 1992). In a novel application of experimental technique, Forsythe et al. (1992) have introduced a computerized field market for candidate-contingent claims to predict the percentage of total vote received by each candidate in an election. Some experimentalists in previous decades tried to measure behavioral parameters or to simulate natural economic processes in the laboratory. For example, Hoggatt (1959) set out to measure oligopolistic "reaction functions," and Garman (1976) tried to simulate the New York Stock Exchange. Experimental economists now recognize that behavioral parameters usually vary with the institution and the environment, so the external validity of such measurements is questionable. As explained in Section 2.1, experimentalists no longer see simulation (in the sense of replicating a field environment as closely as possible) as a useful goal.

A related but more modest purpose for experiments has recently emerged. Aircraft engineers find it useful to study a small-scale model in a "test bed" before trying to build and fly a new plane. Likewise, economists and policymakers recently have found it useful to study new institutions in the laboratory before introducing them in the field. McCabe et al. (1991) describe "test-bed" experiments of computer-aided markets for composite commodities such as computer resources, and gas and electrical power grids. Given the accelerating pace of transformation in the formerly centrally planned economies and given continuing deregulation in Western economies, the scope for institutional engineering of this sort is large and increasing.

Finally, experiments have an important pedagogical purpose. The first recorded use of economics experiments, by Chamberlin (1948), was primarily pedagogical. Since the 1980s this use of economics experiments has grown steadily. Incorporating experimental demonstration of economic propositions into the high school and college curriculum is a natural accompaniment of the evolution of economics as an experimental science. Walker, Williams, and their colleagues at Indiana University, and Wells and his colleagues at the University of Arizona have developed many pedagogical economics experiments (see Wells, 1991; Williams and Walker, 1993).

2

Principles of economics experiments

How do you choose and present the rules governing an experimental economy? How do you choose and motivate subjects? The principles presented in this chapter will provide some guidance. We begin with by discussing the relations between laboratory experiments, formal models, and reality, then informally present the key concepts of economic agents and economic institutions. (See Smith, 1976, 1982b for a more formal presentation using the framework of Hurwicz, 1972.) The next few sections present induced-value theory (again based on Smith, 1976). After a general discussion of external validity or *parallelism*, we highlight some practical implications and apply the ideas to an important strand of literature on market experiments.

2.1 Realism and models

Unless you already are an experienced experimentalist, your first instinct in designing an experiment probably will be to pursue *realism* – design the laboratory environment to resemble as closely as possible a real-world environment of substantive economic interest. If you are interested in securities markets, for example, you might have some subjects serve as investors, some as floor brokers, and some as specialists, all following the rules of the New York Stock Exchange.

On the other hand, if you are a theorist, your first instinct might be to design an experiment that replicates as closely as possible the assumptions of a formal model of interest. For the securities market example, you would throw out the brokers and specialists and the New York Stock Exchange rules, and perhaps ask your subjects to reveal their optimal demand/supply schedules to a (Walrasian) auctioneer.

Which approach is right: to mimic reality or to mimic a formal model? The correct answer is *neither*. Your goal should be to find a design that

offers the best opportunity to learn something useful and to answer the questions that motivate your research. Usually an effective design is quite simple compared to reality, and in some respects simpler than relevant formal models.

It is futile to try to replicate in the laboratory the complexities of a field environment. Like fractals, reality has infinite detail; it is its own best model. No matter where you stop in building the details of reality into your laboratory environment, an infinite amount of detail will always remain uncaptured. A practical difficulty is that your budget probably won't let you get far in this direction. Before you get close, the laboratory environment will have become so complex that you will find it difficult or impossible to disentangle causes and effects. As in any other experimental discipline, simplicity enhances control. Try to find the simplest laboratory environment that incorporates some interesting aspects of the field environment. In the asset market example, to discover whether the market disseminates insider information, you will learn more if you begin with a single, simple, tradeable security and find an appropriate way to feed some traders inside information on its fundamental value.

It is equally futile to try to replicate in the laboratory the precise assumptions of a formal model. A practical difficulty is that most formal models leave out details, and you typically must make choices that are arbitrary in terms of the theory but important in terms of behavior. For example, in a rational expectations model, traders' orders theoretically are based on observed market-clearing prices. In the laboratory, do you announce market-clearing prices before traders place orders or after? Either way you fail to replicate the formal model.

Even if you succeed in creating a laboratory economy that closely replicates the assumptions of a formal model, you usually will not learn much from it. If the observed behavior in your economy is consistent with the implications of the formal model, you have only weak evidence of the model's explanatory power. The evidence would be stronger if you had observed the same behavior in a laboratory economy that relaxed the more stringent assumptions of the model. Suppose you somehow were able to recreate precisely the formal model in the laboratory. Data consistent with the model only tells you that there is no obvious logical flaw in the model – a hollow victory at best, since laboratory experiments are less efficient in detecting logical errors than mathematical analysis or computer simulation. On the other hand, if the observed behavior you report is inconsistent with a logically valid formal model, you face criticism that your design was inadequate or that your subjects failed to understand the environment or both. Unless your

purpose is to demonstrate the model's narrow or empty range of applicability (as in Isaac and Smith, 1985), you can learn rather little from such an exercise.

An analogy may clarify the relationships between reality, formal models, and laboratory experiments. An artist wishes to express a human event, say the death of his brother. He is unable to reenact the real event (that brother is gone) and finds it undesirable for practical and aesthetic reasons (not to mention moral reasons) to replicate it closely. He chooses a medium of expression, perhaps canvas or stone. The quality of his painting will be judged by how well it simplifies reality to capture and communicate the essence of his loss. The stone sculpture also will be judged by its impact on the viewer, not by its fidelity either to reality or to the painting. Likewise, a laboratory experiment should be judged by its impact on our understanding, not by its fidelity either to reality or to a formal model.

2.2 Controlled economic environments

An experiment takes place in a controlled economic environment. Controlled or otherwise, an economic environment consists of individual economic *agents* together with an *institution* through which the agents interact. For example, the agents may be buyers and sellers and the institution may be a particular type of market. Another example, drawn from politics, has voters as agents and majority rule as an institution.

Agents are defined by their economically relevant characteristics: preferences, technology, resource endowments, and information. Your subjects have their own home-grown characteristics, but often you want to examine theories that assume specific characteristics that may or may not correspond to those of available subjects. You might think at first that agents' characteristics are difficult to observe, much less control. The next subsection explains how induced-value theory (Smith, 1976) identifies sufficient conditions for experimental control, conditions that are often easy to satisfy in practice.

An economic institution specifies the actions available to agents and the outcomes that result from each possible combination of agents' actions. Achieving experimental control over the institution is conceptually straightforward: The experimenter explains and enforces the rules. Specific techniques are discussed later.

2.3 Induced-value theory

The key idea in induced-value theory is that proper use of a reward medium allows an experimenter to *induce* prespecified characteristics in experimental subjects, and the subjects' innate characteristics become largely irrelevant.

Three conditions suffice to induce agents' characteristics:

1. *Monotonicity.* Subjects must prefer more reward medium to less, and not become satiated. Formally, if $V(m,z)$ represents the subject's unobservable preferences over the reward medium (m) and everything else (z), then the monotonicity condition is that the partial derivative V_m exists and is positive for every feasible combination (m,z). This condition seems easy to satisfy by using domestic currency as the reward medium.
2. *Salience.* The reward Δm received by the subject depends on her actions (and those of other agents) as defined by institutional rules that she understands. That is, the relation between actions and the reward implements the desired institution, and subjects understand the relation. For example, a $5.00 fixed payment to subjects for participating is not salient because the payment does not depend on the subjects' choice of actions in the laboratory after she shows up. On the other hand, a payment of one cent for every point of profit earned in a market experiment is salient because the payment depends on subjects' actions.
3. *Dominance.* Changes in subjects' utility from the experiment come predominantly from the reward medium and other influences are negligible. This condition is the most problematic of the three since preferences V and "everything else" z may not be observable by the experimenter. Dominance becomes more plausible if the salient rewards Δm are increased and if the more obvious components of z are held constant. For example, subjects often care about the rewards earned by other subjects. If the experimental procedures make it impossible to know or estimate others' rewards (Smith calls this *privacy*) then a component of z is neutralized. Demand effects, arising from subjects' efforts to help (or hinder) the experimenter, are a second example. As the experimenter, avoid revealing your own goals and you neutralize another component of z.

When the three conditions are satisfied, the experimenter achieves control over agents' characteristics. To illustrate, suppose you want to induce some specific smooth preferences (e.g., Cobb-Douglas) represented by the utility function $U(x,y)$. You pick convenient objects such as colored slips of paper, say x = number of slips of red paper and y = same for blue, and clearly explain to the subject (e.g, using

a table of rewards with columns indexed by x and rows by y) that her payment will be $\Delta m = U(x,y)$. Then the induced preferences are $W(x,y) = V(m_0 + U(x,y), z_0 + \Delta z)$, where (m_0, z_0) is the subject's unobservable initial endowment of money and everything else, and Δz summarizes the subject's nonpecuniary proceeds from the experiment. By Hicks's Lemma (1939, appendix) we may conclude that two utility functions represent the same preferences if their marginal rates of substitution always coincide. We have

$$MRS^W = \frac{W_x}{W_y} = \frac{V_m U_x + V_z \Delta z_x}{V_m U_y + V_z \Delta z_y} = \frac{V_m U_x}{V_m U_y} = \frac{U_x}{U_y} = MRS^U$$

with the first and last equalities following from a standard property of marginal rates of substitution, the second equality from salience and the chain rule of calculus, the third equality from (complete) dominance, and the fourth equality from monotonicity. Thus the prespecified preferences represented by U and the induced preferences represented by W are indeed the same.

The intuition is that the experimenter can freely choose any relationship between intrinsically worthless objects and the reward medium. As long as he can explain the relationship clearly to the subjects (salience) and subjects are motivated by the reward medium (monotonicity) and not other influences (dominance), then the experimenter can control subjects' characteristics to implement the chosen relationship in laboratory. A standard example is sellers' cost in a market experiment. If you want to implement increasing marginal costs, say $c_1 < c_2 < c_3$ for three indivisible units, you simply tell the subject that she will receive m units of the reward medium, where $m = (p_1 - c_1) + (p_2 - c_2) + (p_3 - c_3)$, if she sells the units at successive transaction prices p_1, p_2, and p_3.

The concept of salience differentiates surveys from controlled economics experiments. A typical survey asks respondents to report some aspects of their personal characteristics, historical actions, or events. U.S. Bureau of Labor Statistics gathers, classifies, and reports a great deal of such happenstance data. In addition, survey technique is sometimes used to ask respondents to make choices in hypothetical situations. Controlled economics experimentation must not be confused with this latter class of surveys; since no salient rewards are offered in such surveys, respondents are not making economic choices under conditions within the control of the researcher. A laboratory procedure that pays subjects a flat participation fee to respond to hypothetical choices, properly speaking, is a survey and not a controlled economic experiment, because rewards are not salient. (See Kotlikoff, Samuelson, and Johnson 1988, for an example.) What people say they would do in hypothetical situations does not necessarily correspond to what they actually do (see Bishop, 1986). On the other hand, a field "market survey" that offers a choice between brand X and brand Y is a controlled economic experiment if respondents know they get to keep the brand they choose.

We should note that some economics experiments, especially early pilot studies, continue to be conducted and reported without salient rewards. Sometimes salient rewards substantially alter the experimental outcomes and sometimes they don't; Jamal and Sunder (1991) find that the use of salient rewards tends to increase the reliability of results; see Smith and Walker (1992) for a recent summary of the evidence. In any case, an experimentalist who uses unmotivated subjects can anticipate that many economists will challenge the results.

2.4 Parallelism

Some economists question the external validity of laboratory data and feel that such data somehow is not representative of the real world. For example, in 1987 an anonymous referee of a paper on laboratory asset markets discounted the relevance of the work on the grounds that "experienced traders used to dealing with large sums of money [may not] use the same heuristics, etc., exhibited by rather naive students who may or may not take this seriously." Bohm raises the issue in motivating his field experiment: "If a given mechanism can be shown to work . . . in one, two or three laboratory tests, how can we be sure it will work in the fourth instance when we want an important decision to be determined by it?" (1984, p. 137).

Experimentalists in other disciplines have encountered similar skepticism. Galileo's critics did not believe that the motion of pendulums or balls on inclined planes had any relation to planetary motion in the celestial sphere. More recently, some people question whether substances found to be toxic in large doses for laboratory rats will harm human beings exposed to small doses over longer periods of time.

Deductive logic does not provide the basis to reject such skepticism. From the mere fact that you have observed the sun rise every morning for twenty years you can't really deduce the proposition that it will rise again tomorrow morning. Yet people do make the leap of faith that the sun will rise. This is *induction.*

The general principle of induction is that behavioral regularities will persist in new situations as long as the relevant underlying conditions remain substantially unchanged. Theory suggests what is "relevant" and what is a "substantial" change, but the principle itself is an assumption

(an "axiom" or "maintained hypothesis," if you prefer), not a deducible proposition.

Vernon Smith refers to the induction principle in the present context as the "parallelism precept":

> Propositions about the behavior of individuals and the performance of institutions that have been tested in laboratory microeconomies apply also to nonlaboratory microeconomies where similar *ceteris paribus* conditions hold. (1982b, p. 936)

According to parallelism, it should be *presumed* that results carry over to the world outside the laboratory. An honest skeptic then has the burden of stating what is different about the outside world that might change results observed in the laboratory. Usually new experiments can be designed and conducted to test the skeptic's statement. For example, in the past both authors have heard colleagues argue that laboratory asset market data are "artificial." When pressed, the colleague usually cites the large number of traders or the high stakes and the professionalism of traders in the real world as the important differences. The appropriate response is to conduct experiments with more traders or more experienced (or professional) traders or to increase the salient rewards. The idea is to use the skepticism to promote constructive research, and not to engage in sterile arguments.

For scientific purposes, the simplicity and small scale of laboratory environments relative to field environments are virtues. Charles Plott makes the case as follows.

> The art of posing questions rests on an ability to make the study of simple special cases relevant to an understanding of the complex. General theories and models by definition apply to all special cases. Therefore, general theories and models should be expected to work in the special cases of laboratory markets. As models fail to capture what is observed in the special cases, they can be modified or rejected in light of experience. The relevance of experimental methods is thereby established. (1982, p. 1520)

In the same article, Plott deals with general concerns regarding external validity as follows:

> While laboratory processes are simple in comparison to naturally occurring processes, they are real processes in the sense that real people participate for real and substantial profits and follow real rules in doing so. It is precisely because they are real that they are interesting. (p. 1486)

2.5 Practical implications

A few minutes' reflection on induced-value theory yields some basic practical advice for beginners on the conduct of economic experiments. Among the more important do's and don't's:

1. To create controlled economic environments in laboratory, motivate subjects by paying them in cash. (Grades may also work for student subjects; see Chapter 4). Most of the payment should be sensitively linked to subjects' actions in the experiment. The average payment should exceed subjects' average opportunity cost. Such payments promote monotonicity and salience.
2. Find subjects whose opportunity costs are low and whose learning curves are steep, in order to achieve dominance and salience at moderate cost. Undergraduate students are usually a good bet.
3. Create the simplest possible economic environment in which you can address your issues. Simplicity promotes salience and reduces ambiguities in interpreting your results. Check instructions carefully for accuracy and clarity. Verify subjects' understanding in "dry runs" or quizzes.
4. To promote dominance, avoid loaded words in instructions. In a prisoner's dilemma experiment, for example, label the choices A and B rather than Loyal and Betray. Use neutral terms for subjects' roles – for example, buyer and seller or player A and player B rather than czar and serf or opponent.
5. If dominance becomes questionable and your budget permits, try a proportional increase in rewards. A systematic change in observed outcomes suggests that dominance had not been achieved at the lower level of rewards.
6. When feasible and appropriate for your research, maintain the privacy of subjects' actions and payoffs, and of your own experimental goals. Subjects' homegrown (i.e., innate) preferences may have rank-sensitive malevolent or benevolent components that will compromise dominance when privacy is not maintained.
7. Do not deceive subjects or lie to them. It is true that social psychologists have sometimes run interesting experiments based on deception (e.g., Stanley Milgram, 1974). However, experimental economists require complete credibility because salience and dominance are lost if subjects doubt the announced relation between actions and rewards, or if subjects hedge against possible tricks. Deception harms your own credibility and that of

other experimentalists, thereby undermining the ability to achieve experimental control.

These rules are not ironclad. For example, there are advantages to using unpaid subjects in early pilot experiments. Later chapters will delve more deeply into the art of writing instructions, the circumstances in which privacy is appropriate, and so on. We suggest that you feel free to break these rules, but only when you are confident that you understand the underlying issues and that you can convince most skeptics that your reasons are sufficient.

2.6 **Application: The Hayek hypothesis**

The efficiency of competitive equilibrium (CE), popularly known as Adam Smith's Invisible Hand Theorem, is universally acknowledged as a central proposition in economics. However, economists differ sharply on the conditions necessary for the attainment of CE and therefore on the practical significance of the proposition. The usual textbook explanation, and perhaps the majority view among economists, is that the conditions are quite stringent, including (a) large numbers of buyers and sellers, each small relative to the market, who possess (b) perfect or at least very good information about demand and supply conditions. Other economists, an influential minority, believe the proposition holds given only a moderate number of buyers and sellers with little or no public information other than current prices. Friedrich Hayek, for example, states:

> The most significant fact about this [price] system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take the right action. (1945, pp. 526–7)

Edward Chamberlin, an influential proponent of the majority view, addressed this range-of-applicability controversy in one of the earliest laboratory studies in economics. He created a simple classroom environment that incorporated what he viewed as key aspects of ongoing field markets: fairly large numbers of transactors (dozens) with imperfect information and no central auctioneer to coordinate trade. Chamberlin assigned (as private information) single unit values and costs to students who acted as buyers and sellers. The sellers and buyers searched for counterparties and set transaction prices in bilateral negotiations. Chamberlin reported considerable dispersion and some bias in transaction prices and significant inefficiency, due mostly to transactions involving either an extramarginal buyer or an extramarginal seller. He concluded:

> My own skepticism as to why actual prices should in any literal sense tend toward equilibrium during the course of a market has been increased not so much by the actual data of the experiment before us – which are certainly open to limitations – as by failure, upon reflection stimulated by the problem, to find any reason why it should be so. It would appear that, in asserting such a tendency, economists may have been led unconsciously to share their unique knowledge of the equilibrium point with their theoretical creatures, the buyers and sellers, who, of course, in real life have no knowledge of it whatever. (1948, p 102)

Vernón Smith (1962) reported another set of simple laboratory markets based on a different view of the important aspects of ongoing field markets. Like Chamberlin, he used dozens of undergraduate buyers and sellers with privately assigned values and costs, but changed the laboratory environment in two important respects. Smith employed the double-auction (DA) institution in which buyers and sellers transact by making and accepting public bids and asks, rather than Chamberlin's bilateral search institution. Smith also used stationary repetition, in which value and cost assignments are held constant across several trading periods. He found that transaction prices converged reliably and fairly quickly to CE values. Plott and Smith (1978) discovered that the efficiency of such markets was always quite high, often 100 percent.

Thousands of experiments since then have corroborated Smith's results. Indeed, only a few buyers and sellers (two to four each) are required to achieve rapid convergence to efficient CE outcomes when subjects are paid according to the precepts of induced-value theory. Smith summarizes the findings in terms of what he calls the "Hayek Hypothesis: Strict privacy [regarding agents' value and cost characteristics] together with the trading rules of a market institution are sufficient to produce competitive market outcomes at or near 100% efficiency" (1982a, p. 167). The evidence strongly supports the hypothesis in simple stationary-repetitive environments using the DA institution. More complex laboratory environments using several alternative market institutions also generally support the hypothesis (but see Holt, Langan, and Villamil, 1986, and Davis and Williams, 1991, for some qualifications). Smith exercises caution in interpreting the findings:

> What has been established is, that in the simple environments studied to date, the attainment of C. E. outcomes is possible under much less stringent conditions than has been thought necessary by the overwhelming majority of professional econ-

> omists. . . . But even if our Hayek hypothesis continues to outperform its competitors in laboratory experiments, does this mean it will do comparably well in the "field" environment of the economy? On the assumption of parallelism, namely that the same physical (and behavioral) laws hold everywhere, it is a reasonable working hypothesis, provisionally, to make this extension, but independent field observations, or experiments, are the appropriate vehicle for testing the extended hypothesis. (1982a, p. 177).

Gode and Sunder (1992, 1993a,b) illustrate the fruitful interplay between experiment and computer simulation, and add a new twist on the Hayek hypothesis. The authors create zero-intelligence (ZI) computerized traders that bid or ask randomly subject to a no-loss constraint. They find that the double-auction institution produces highly efficient outcomes even with ZI traders! Perhaps the rationality assumption plays a smaller role in some market institutions than most economists have presumed.

3

Experimental design

How does the number of buyers and sellers affect market efficiency? Do consumers prefer the "new improved" product or the "classic" version? Whether your purposes are scientific or commercial, you probably are interested in the effects of only a few variables, the *focus* variables. Usually you must also keep track of several other variables of little or no direct interest, the *nuisance* variables, because they may affect your results.

Which variables are focus and which are nuisance in your experiment depends on your purpose. The number of buyers is a focus variable in some oligopoly experiments, but the same variable is a nuisance in experiments testing consumer response to new products.

This chapter will explain how to design experiments that sharpen the effects of focus variables and minimize blurring due to nuisance variables. It will also explain how to design experiments that allow you to disentangle the effects of different variables, that is, how to avoid *confounding* the effects of two or more variables.

The first two sections introduce control and randomization, the basic ingredients of proper experimental design. Sections 3.3 and 3.4 elaborate on these ingredients and discuss specific designs. Distilled practical advice appears in the next section, and the last section illustrates the main ideas while reviewing some "test-bed" market experiments.

A word of warning before we begin. This chapter contains technical jargon. We have tried to follow the most common practices, but the literature is not entirely consistent in how words are used. You can consult the glossary at the end of the book to see how we use these words, but be careful in reading the literature to check what the author really means.

3.1 Direct experimental control: Constants and treatments

In the laboratory you can directly control many variables. You can freely select cost and value parameters and trading rules in market experiments, or the choice set and the subject pool in individual choice experiments. By controlling important variables you produce experimental data rather than happenstance data.

The simplest way to control a variable is to hold it *constant* at some convenient level. For example, enforce the same double-auction trading rules throughout a market experiment. The main alternative is to chose two or more different levels that may produce sharply different outcomes, and to control the variable at each chosen level for part of the experiment (or subset of experiments). For example, use two different sets of cost parameters, one inducing highly elastic supply and the other inelastic supply. Perhaps because of their prevalence in medical experiments, variables controlled at two or more levels are called *treatment* variables.

There is a tradeoff between controlling variables as constants and as treatments. As you hold more variables constant your experiment becomes simpler and cheaper, but you learn less about the direct effects and the interactions among the variables. Section 3.5 offers some suggestions on managing this tradeoff.

Suppose you choose two treatment variables, say the market institution with levels PO (posted offer) and DA (double auction), and the demand elasticity with levels E (elastic) and I (inelastic). Despite your control, you will completely confound their effects if you always change the variables together, say PO-E combination half the time and DA-I combination the other half. Instead, if you run each treatment combination (PO-E, PO-I, DA-E, and DA-I) one quarter of the time, you can gauge the separate effects of the two treatments. The logic is quite general: *Vary all treatment variables independently* to obtain the clearest possible evidence on their effects (see Figure 3.1).

3.2 Indirect control: Randomization

Some variables are difficult or impossible to control. For example, weather is an important and uncontrollable nuisance in agricultural experiments. (And occasionally in economic experiments: One of the authors recalls snowstorms preventing subjects from showing up and the other author remembers watching helplessly as airconditioning failed and the room temperature rose above 100°F in an early computerized experiment.) For economists, subjects' expectations usually are more important than the weather and just as uncontrollable. Some potentially

A. Confounded Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	No Observations
Double Auction	No Observations	Observations (DA-I)

B. Independent Treatment Variables:

	Elastic Demand	Inelastic Demand
Posted Offer Auction	Observations (PO-E)	Observations (PO-I)
Double Auction	Observations (DA-E)	Observations (DA-I)

Fig. 3.1 Independent variation of treatment variables.

important nuisances, such as a subject's alertness and interest, are not even *observable* by the experimenter, much less controllable.

Uncontrolled nuisances can cause inferential errors if they are confounded with focus variables. The real cause of improvement in harvests in the year a new seed variety is introduced may be good weather. Efficiency may decline when elastic supply parameters are introduced late in a long experiment, but the reason may be subjects' fatigue. The problem is that you may attribute an observed effect to a focus variable when the effect actually arises from an uncontrolled nuisance.

How can you avoid confounding problems when you can't directly control some important nuisances? The advice offered at the end of the previous section provides a hint. Independence among controlled var-

iables prevents confounding problems. We would solve the present problem if we could somehow make the uncontrolled nuisances independent of the treatment variables.

Randomization provides indirect control of uncontrolled (even unobservable) variables by ensuring their *eventual* independence of treatment variables. The basic idea is to assign chosen levels of the treatment variables in random order. For example, in a market experiment subjects' personal idiosyncracies and habits are an uncontrollable and largely unobservable nuisance variable. When subjects arrive, don't assign all the early birds to the role of sellers and the late arrivals to the role of buyers. Randomize the assignment and you can be confident that observed profit differences between buyers and sellers arise from differences in the roles and not from differences in subjects' personal characteristics.

The simplest valid experimental design is called *completely randomized.* In this design, each treatment (or each conjunction of treatment variables) is equally likely to be assigned in each trial. (A *trial* is an indivisible unit of an experiment, such as a trading period in a market experiment.) Suppose you choose a completely randomized design for the two-treatment experiment illustrated in Figure 3.1. Then in each trial you might flip two fair coins to select each of the four treatments PO-E, PO-I, DA-E, and DA-I with probability 0.25 in each trial, independently of selections in previous trials.

Complete randomization is quite effective when you can afford to run many trials. Independence among your treatment variables and uncontrolled nuisance variables is "eventual" in the sense that only as the number of trials gets arbitrarily large does the probability of a given positive or negative correlation go to zero. You can occasionally get a large correlation between treatments and uncontrolled nuisances in a small set of randomized trials. Classical statistical techniques, discussed in Chapter 7, take this problem into account.

When uncontrolled nuisances produce little variation across trials, the completely randomized design is hard to improve upon. When controllable nuisances do significantly affect outcomes, however, designs that appropriately combine control with randomization are more efficient in the sense that they can produce equally decisive results from fewer trials. These designs ensure zero correlation among controlled variables even in small sets of trials.

Random block is the general name given to this improved design. The difference from the completely randomized design is that one or more nuisance variables are controlled as treatments rather than randomized.

Nuisance treatment variables are often called blocking variables, held constant within a block [subset of trials] but varied across blocks. The next two subsections provide examples.

3.3 The within-subjects design as an example of blocking and randomization

The purpose of the classic boys' shoe experiment (Box, Hunter, and Hunter, 1978, p. 97ff) is to see whether a new sole material lasts longer than the old. The focus is sole material, a treatment variable with two levels: old and new. Measured wear varies considerably, mostly from subjects' different activities and habits: Some boys are couch potatoes, others ride scooters using a shoe for a brake. Clever experimental design prevents these nuisances from obscuring the focus variable's effects: Each boy gets a pair of shoes with one sole of new material and the other sole of old. Thus subject identity in this design is a blocking (i.e., nuisance treatment) variable that captures the habits and activities nuisances, and *differences* in measured wear between left and right soles becomes the relevant performance measure. Random assignment of the focus variable (new material on left or right shoe) reduces confounding due to other nuisances, such as whether scooter brakers tend to be left or right footed.

Experimental designs that vary levels of the focus variable only across subjects are generically called *between subjects* designs and those that use several different levels for each subject are called *within-subjects* designs. The shoe experiment uses a special within-subjects design that allows all data to be expressed as differences across matched pairs. The matched-pair differences allow sharper inferences to the extent that individual subject variation is an important nuisance.

The same trick can be useful in economics experiments. For example, suppose you conduct individual choice experiments comparing the willingness to pay (WTP) for a gamble to the willingness to accept (WTA) a certain payment in lieu of the gamble. If you want to see whether your new "transparent" instructions will bring WTP and WTA closer together, then individual variability is an important nuisance you should take into account – for instance, some subjects may be more risk averse than others and report low WTP and low WTA. It would be appropriate to employ a within-subjects design as in the shoe experiment. Specifically, you could ask each subject for WTPs and WTAs in random order, and analyze the *differences* WTA − WTP across subjects for each gamble. That way you eliminate a potentially important source of noise, and the effects of your focus (instructions) then become more visible.

3.4 Other efficient designs

The within-subjects idea has two useful variants. A *crossover design* takes a subject or group of subjects and varies the levels, say A and B, of a treatment variable across trials. When you suspect your treatment variable has effects lasting several trials, you should consider the ABA crossover design. (The simpler AB design confounds time and learning with the treatment variable.) For example, suppose your focus variable is the market institution with A = the double auction and B = buyers' auction (sellers passive). The convergence behavior of a group of traders may carry over from one trading period to the next, so in one session you might conduct four A trading periods followed by eight B trading periods and finish with four more A periods (ABA), and use the complementary BAB design in a companion session. Then the difference in mean observed performance between the A and B periods would conservatively indicate the effect of your focus variable.

A second variant, the *dual trial*, is especially useful when individual or group idiosyncrasies may be an important nuisance. Kagel and Levin (1986), for example, suspected that individual random signals and the behavior of other bidders in a group could affect bidder behavior in first-price common-values auctions. To test cleanly the effects of the focus variable, group size with levels *S*(mall) and *L*(arge), they employed dual auctions: upon receiving her signal, each subject submitted two bids, one for a small-group auction and a second for the large-group auction. Their dual auction design allowed the authors to isolate the effect of group size by looking at differences $(b_L - b_S)$ in the two bids across subjects and time periods.

The *factorial design* is perhaps the most important general method for combining randomization and direct control when you have two or more treatment variables. To illustrate, consider two treatment variables ("factors") labeled *R* and *S*, with three levels H(igh), M(edium) and L(ow) for *R* and two levels H(igh) and L(ow) for *S*. In the resulting 3 × 2 factorial design, each of the six treatments LL, LH, ML, MH, HL, and HH is employed in the same number *k* of trials. Thus $3 \times 2 \times 4 = 24$ trials are required to replicate the design $k = 4$ times. Randomization plays an essential role in that you must assign the six treatments in random order to the six trials in each replication.

When it is feasible, the factorial design is more efficient than the completely randomized design because it ensures that each treatment (combination) occurs an equal number *k* of times, and that the treatment variables all have zero correlations even for small replication numbers *k*. Among other things, this helps you to distinguish the direct effects of the treatment variables from interactions.

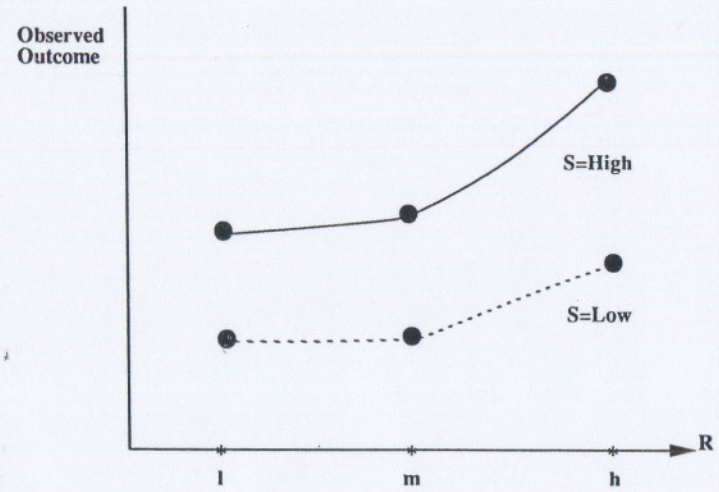

Fig. 3.2 Mean outcomes in a hypothetical factorial experiment.

Figure 3.2 uses the 3×2 example to illustrate direct and interactive effects. The vertical axis is the observed outcome, say market efficiency. The first treatment variable *R*, say elasticity of demand and supply, appears on the horizontal axis and the second variable *S*, say payoff intensity, shows up in the two curves labeled S = High and S = Low. The curves themselves connect the hypothetical mean outcomes in each treatment. The distance between the curves measures the direct effect of variable *S*. When the curves are parallel, there is no interaction between *R* and *S*, but when the gap between the curves widens as in Figure 3.1, there is a positive *RS* interaction. Chapter 7 will discuss the issue more extensively.

The factorial design is a bit less robust than the fully randomized design because experimenter errors in assigning treatments and missing trials (from computer glitches or no-show subjects, for instance) more seriously impair the data analysis. Indeed, if these problems are frequent, the factorial design becomes indistinguishable from the completely randomized.

Another problem with the basic factorial design is that the number of required trials increases quickly as the number of factors increases. Suppose, for example, you chose only two levels for each treatment variable. Even then, you need $2^4 = 16$ trials for 4 factors and $2^8 = 256$ trials for eight factors to run just a single replication! The problem is

serious because there are many potentially important nuisance variables in some economic environments.

The fractional factorial design alleviates the problem. The basic idea is to run a balanced subset of the factorial design. To take the simplest example, suppose you have three variables, each with two levels denoted + and −, and can conduct only four trials. That is, you can run only half of the eight possible treatments (+ + +, + + −, + − +, + − −, − + +, − + −, − − +, and − − −). Your first thought might be just to run the first four treatments on the list, or every other treatment, but a moment's reflection shows that these choices are unbalanced because some variables are held constant or some pairs of variables are correlated. You get a balanced subset of treatments if you impose the restriction that the third sign is the product of the first two. Then the subset of treatments you run is + + +, + − −, − + −, and − − +. If you run this subset (in random order, of course!), then you have a half factorial $2 \times 2 \times 2$ design. If you are a geometric thinker, you can visualize the balance of this design by thinking of each possible treatment combination as a corner of the unit cube in the space of the three treatment variables. For example, + + − could label the upper left back corner and − − + label the lower right front corner. The chosen treatments' center of mass is the center of the cube, and the center of mass on each face is the center of the face. Each level of each treatment variable appears in the same number of trials (2) and each pair of treatment variables is orthogonal.

Conceptually (although not visually) it is straightforward to generalize to more treatment variables and to smaller replication fractions. For example, Copeland and Friedman (1987) use a half-factorial $2 \times 2 \times 2 \times 2$ design in an asset-market experiment, where the fourth treatment variable (infocontent, a focus variable that defines the informational complexity of the environment) is constrained to be the product of the first three treatment variables (two nuisance variables called learnops and paymethod and another focus variable called infoarrival). A more dramatic example is given by Box et al. (1978, p. 394). They present a 2^7 sixteenth-factorial design for determining which of seven variables (seat position, handlebar position, tire pressure, etc.) affect a bicyclist's performance. Only 8 trials are required, compared to 128 in the full once-replicated factorial.

The elegance and economy of the fractional factorial design come at a price. The design obviously is less robust than a randomized design; it loses appeal if you are not confident of your ability to conduct all trials flawlessly. (If you are confident, the design has a subtle advantage: You can complete the factorial design if it turns out you can run additional trials.) The other disadvantage is inherent in the design. The fractional factorial achieves balance in a subset of the possible treatments by systematically confounding some direct effects with some interactions. The simple half-factorial $2 \times 2 \times 2$ example confounds the third variable with the pairwise interaction of the first two variables, for instance. This disadvantage is not always serious. If you know that some pairwise or higher-order interactions are negligible, then you can harmlessly confound them.

We close this section with some background information for readers who wish to learn more about classical experimental design. R. A. Fisher and his colleagues developed most of the concepts presented in this chapter between 1910 and 1940. Much of the terminology comes from agricultural experiments; blocks, for example, originally referred to adjacent rectangular pieces of land, and a split-plot design (a type of randomized block) originally involved subdividing such a block for one treatment variable.

Statisticians with a combinatorial bent noticed that further efficiency gains theoretically arise from imposing additional symmetries on block and factorial designs. For instance, in testing four tire brands (*a, b, c,* and *d*) using four test cars, you could require not only the ordinary blocking condition that each car uses each brand, but also balance the assignment of tires to the four wheels of the test cars – say, use the order *abcd* for the four wheels in the first car, *dabc* in the second, *cdab* in the third, and *bcda* in the fourth car. This design is called Latin square after its diagrammatic representation, and it has higher-dimensional analogues called Graeco-Latin and hyper-Graeco-Latin designs. Such constructions quickly become quite Baroque and are not at all robust to missing trials and so forth.

The interested reader can find dozens of advanced books on experimental design, mostly of the 1950–70 vintage, in the QA279 section (under the Library of Congress system) and other sections of a good library. In writing this chapter we relied most heavily on Box et al., (1978) as well as Campbell and Stanley (1966), and Kirk (1982).

3.5 Practical advice

Theoretical considerations regarding experimental design do have practical consequences. Drawing on the theory, we offer some general advice regarding typical nuisance variables, the choice of constant and treatment variables, and the general conduct of experiments.

3.5.1 Chronic nuisances

Remember that the distinction between nuisance and focus variables depends on your purpose. Experience and learning, for example,

are nuisances if you want to test a static theory but are focus variables if you want to characterize behavioral change over time. This chapter has already mentioned most of the important nuisance variables you typically face in conducting an economics experiment, and suggested ways for dealing with them. Chapters 4 and 7 provide a more systematic discussion, but a quick summary may be useful at this point.

1. Experience and learning: Subjects' behavior changes over time as they come to better understand the laboratory environment. When this is a nuisance, control it as a constant by using only experienced subjects, or control it as a treatment (blocking variable) by using a balanced switchover design.
2. Noninstitutional interactions: Subjects' behavior may be affected by interactions outside the laboratory institution. For example, sellers may get together during a break and agree to maintain high prices. Careful monitoring during the break, or a change in parameters after the break, therefore may be advisable.
3. Fatigue and boredom: Subjects' behavior may change over time simply as a result of boredom or fatigue. For example, after playing strategy A for 58 periods in a repeated prisoner's dilemma, a subject may choose strategy B (defect) just to relieve the tedium. We recommend occasional payoff switchovers and planned sessions of at most two hours for most experiments.
4. Selection biases: The subjects or their behavior may be unrepresentative because their selection was biased. For example, self-selection may upwardly bias self-reported sexual activity when only the most talkative choose to respond to your questionnaire. Experimenter selection may be biased when students in an advanced finance class are recruited for an asset-market experiment. Recognizing the problem is the key step in finding ways to deal with selection biases.
5. Subject or group idiosyncrasies: A subject's background or temperament may lead to unrepresentative behavior. A group of subjects somehow may reinforce each other in unusual behavior patterns. Replication with different subjects therefore is essential.

3.5.2 Disposition of variables

We offer the following suggestions on choosing treatment and constant variables.

1. Control all controllable variables. Otherwise your data will be less informative than they could be.
2. Control focus variables as treatments. Use widely separated levels to sharpen the contrasts. Use two levels and skip intermediate levels unless you are interested in possibly nonlinear effects.
3. When you suspect that a nuisance variable interacts with a focus variable, consider controlling the nuisance as a treatment. Two levels often suffice.
4. Control most nuisances as constants to keep down complexity and cost. Even a nuisance with large effects can harmlessly be held constant as long as its effects are independent of the focus variables' effects.
5. Vary your treatments independently to maximize the resolution power of your data and to avoid confounding.

3.5.3 Phases of experimentation

A laboratory investigation typically proceeds in phases. The preliminary phase identifies the specific issues to be investigated and the essential aspects of the laboratory environment. The next phase consists of one or more pilot experiments. Here you complete the specification of the laboratory environment, prepare instructions for subjects, and conduct the pilot experiments, perhaps with unpaid subjects at first. The results usually lead to improving (simplifying) the instructions and the environment. At this point you should choose the focus and important nuisance variables you will use as treatments; the suggestions in the previous subsection may help.

Now you are ready to begin the formal part of your research by conducting a set of exploratory experiments. You should pick a simple design capable of detecting gross effects of the treatment variables, perhaps a fractional factorial or a $k = 1$ factorial. When you analyze the data you may decide to hold constant some variables that seem to have no interesting effects or interactions. Possibly you will want to adjust the environment or introduce a new treatment variable on the basis of the exploratory data. If you are exploring a new area, you may well discover at this point that major changes in instructions or treatments are necessary. If so, you will probably relabel your work so far as preliminary, and try the second phase again.

The final phase consists of follow-up experiments intended to provide definitive evidence on your chosen issues. Try to reserve 50 to 75 percent of your budget for this phase. If the results of the exploratory experiments seem clear-cut, you may choose simply to replicate them in the

follow-up phase. If the exploratory experiments suggest subtle but relevant direct effects or interactions among your variables, you may choose a more elaborate design.

A final piece of advice. Don't get too fancy in designing your experiments, especially in your first project. Begin with a proven design from related previous research by other authors, or use a simple version of one of the designs we have presented.

3.6 Application: New market institutions

We live in an era of rapid change in economic institutions. Existing markets have expanded and changed, and new markets have opened, in response to advances in computer and telecommunications technology and in response to political developments in Asia, and in Eastern as well as Western Europe. Even in the relatively stable markets of the United States, scandals and technological developments have spurred efforts to reform the primary market for U.S. government securities and the commodity exchanges.

How do we evaluate alternative market institutions? What kinds of market institutions will best promote efficient exchange in the new environments around the world? Existing economic theory and historical experience provide precious little guidance. Field experiments can be costly, as well as politically risky. Laboratory experiments can conveniently serve as test beds for new market institutions. New institutions can be tried out and refined in the laboratory before they are further tested and implemented in the field. This section discusses some of the test-bed work done so far and uses it to illustrate some of the basic principles and issues in experimental design.

> Laboratory experimentation can facilitate the interplay between the evaluation and modification of proposed new exchange institutions before field implementation. . . . Laboratory experiments allow one to investigate the incentive and performance properties of alternative exchange institutions, and, with respect to institutional design, they provide a low-cost means of trying, failing, altering, trying, etc. This process uses theory, loose conjecture, intuitions about procedural matters and, most important, repeat testing to understand and improve the features of the institutional rules being examined. (McCabe, Rassenti, and Smith, 1993, p. 309)

Two kinds of work are discernible in test-bed research. When the institutions are reasonably well-specified, an experiment can be designed using classical approaches discussed in this chapter in order to measure

and compare their performance characteristics. The studies by Hong and Plott (1982) and by Grether and Plott (1984) described below fall into this *performance testing* branch of test-bed research. On the other hand, when the institution itself has to be designed through an iterative design-test-revise process, classical experimental design techniques usually cannot be applied to the overall process, although they may be useful for some phases of the project. This second branch, *developmental testing* is exemplified in Grether, Isaac, and Plott (1981), Plott and Porter (1989), the McCabe et al. (1993) effort to develop a uniform-price double auction, and the McCabe et al. (1988) effort to develop a "smart" market for natural gas. We shall now briefly touch on both branches of test-bed research.

3.6.1 Performance testing

Grether and Plott (1984) conducted some early test-bed experiments dealing with a controversy about existing market institutions. In May 1979 the U.S. Federal Trade Commission filed an antitrust suit against the four domestic producers of a gasoline additive, tetraethyl lead. The suit claimed that uncompetitive high prices were sustained by three institutional practices: advanced notification of price changes (AN), "most favored nation" guarantees to customers that nobody else will get a lower price (MFN), and "delivered pricing" quotes that include transportation cost (DP). The four lead producers argued that the institutional practices were a convenience to customers and had no anticompetitive effects.

In their laboratory study, Grether and Plott break the AN institution down into three focus variables: price publication with three levels (N = no seller publishes prices, L = the two largest sellers publish, and A = all sellers publish prices), price access with two levels (B = only buyers see published prices, and A = all buyers and all sellers see published prices), and advanced notice per se with two levels (Y = yes, a seller can change price only if it is announced in the previous period, and N = no advanced notice required). They made MFN a single two-level (Y or N) variable and omitted DP from their study. Even so, there are potentially $3 \times 2 \times 2 \times 2 = 24$ institutional treatments (i.e, conjunctions of the four treatment variables).

In order to keep the study within budget, Grether and Plott held constant most other relevant variables including supply–demand parameters (at a level chosen to resemble the field conditions) and the basic exchange institution (bilateral search using telephones). Some conjunctions of treatments are vacuous or uninteresting (e.g., access to prices when no sellers publish prices) and some are especially interesting (e.g.,

AAYY = all disputed practices present, and N-NN = all disputed practices absent). Given the time and budget limitations, Grether and Plott used only 8 of the 24 possible treatments in their 11 laboratory sessions of 16 to 25 periods each. The most interesting treatments were used most often and most sessions use an ABA crossover design.

The data clearly support the conclusion that transaction prices are near competitive equilibrium when the disputed practices are absent (e.g., in treatment N-NN) but are substantially higher when the practices are all present in treatment AAYY.

The authors are cautious about drawing firm conclusions for the U.S. lead additive industry. However, they do convincingly argue that the disputed practices could no longer be presumed to be benign. After the experiments, the defendants lost the case to the government in trial but won on appeal. We conclude that the experimental design was adequate for the authors' purposes and that it provides an example of good exploratory work. A more careful design would be necessary in follow-up work to assess the separate and interactive effects of the institutional practices.

An institutional performance test by Hong and Plott (1982) used an even simpler experimental design. Railroad companies lobbied with the Interstate Commerce Commission to require barges to post rates. Railroads argued that publicly posted rates will make the industry more competitive, and protect the smaller barge companies from being secretly undersold by their larger rivals. While railroads had been required to post prices, the dry bulk cargo market on Mississippi operated largely by telephone between carriers and shippers.

Hong and Plott's (1982) simple design had one treatment variable, market organization, that took two values, posted price and telephone market. Two replications required a total of four market sessions. Identical parameters, based on scaled-down judgments of people in the industry, were used in all four sessions. Posted price markets revealed higher prices, lower efficiencies, and lower profits for smaller sellers. The railroads soon backed down from their efforts to change the prevalent rules for the barge market.

3.6.2 Development testing

Developmental test-bed studies are essentially sequential in nature. Since the design of the institution is being evolved, the factorial and other classical experimental designs described in the preceding sections in this chapter cannot be used to structure the overall study, but the general principles of control and randomization remain as important as ever. In the following paragraphs, we give a few examples of developmental test-bedding.

From 1968 through the mid-1970s, landing rights at major U.S. airports (Washington National, Kennedy, La Guardia, and O'Hare) were allocated among airlines by committees consisting of airlines that had been certified by the Civil Aeronautics Board. With the Airline Deregulation Act of 1976, the possibility that these committees could be used as a barrier to new competition arose. To what extent was the committee process, already in place, compatible with the Airline Deregulation Act?

Grether et al. (1981) conducted demonstration experiments with two kinds of institutions, committees and markets. The primary purpose of this experiment was to demonstrate the consequences of alternative decision-making processes. The authors found that (1) the outcome of the committee process is sensitive to the consequences of the default option resorted to in case of a deadlock in the committee; (2) separate committees for different airports could not efficiently handle the interdependencies between the airports; (3) the committee process is insensitive to the profitability of the individual airlines. In the market experiment they found that (1) speculation in landing slots was not a serious problem; (2) price of landing slots was determined not by their value to large airlines but by their marginal value; and (3) market processes can be designed to efficiently solve certain problems that are not solved efficiently by the committee process. Over the years, airlines have come to favor a market process for allocation of airport landing slots though the Federal Aviation Administration favors an administrative solution.

The U.S. Federal Energy Regulatory Commission funded a series of studies on electric power and natural gas networks (see Alger, O'Neill, and Toman, 1987a, b; Alger, 1988; McCabe et al. 1988; and Plott, 1988). As explained in the *Science* magazine overview, "Smart Computer-Assisted Markets," by McCabe et al. (1991), technological progress now allows markets to be created for goods with important indivisibilities and complementarities. For example, a gas distributor will want to make a purchase from a gas producer only if she can also purchase adequate transmission rights from pipeline owners at sufficiently favorable prices. Existing networks and computerized market programs could support the new markets, which promise substantial efficiency gains over traditional contracting arrangements.

For example, price dispersion disrupts markets for highly complementary goods like gas and gas transmission. Despite its great virtues, the double-auction market institution produces dispersed transaction prices, but some alternative market institutions do not. The call (or clearinghouse) institution, for instance, collects all bids and asks during a trading period, aggregates them respectively into demand and supply

curves, and clears the market at a single, uniform price defined by the intersection of supply and demand. For use in markets with the complementary goods, McCabe et al. (1993) design a new market institution, the uniform-price double auction (UPDA) to combine the continuous feedback of the DA with the uniform pricing of the call market. The basic idea (independently explored in Friedman, 1993) is to continually announce the tentative clearing price as bids and asks accumulated during a call market trading period.

McCabe et al. (1993) study 8 variants of UPDA defined by three two-level variables: the call rule (exogenous end to the period at a prespecified time, or endogenous end when some condition holds, say when no new orders arrive for 20 seconds), the update rule (1s or 2s, the distinction involving how much a trader must improve previous offers to transact), and the inform rule (open book = all traders see all tentatively accepted and tentatively rejected bids and asks, and closed book = each trader sees only her own tentatively accepted bids or asks). The authors lay out a $2\times2\times2$ factorial design with eight replications; the design calls for each UPDA variant to be tested in three sessions using subjects experienced in one of the previous five sessions using that variant. The environment is held constant across sessions; it features a supply–demand configuration that shifts up and down randomly from one market period to the next. The authors find that inexperienced subjects do best with the exogenous close, 1s, closed-book variant, and experienced subjects do best with the endogenous close, 1s, open-book variant, and that efficiencies approach those of the basic double-auction market institution.

McCabe et al. provide a good example of first-stage follow-up experiments, given a large budget. Subsequent follow-up experiments will presumably match the best versions of the UPDA institution against other promising market institutions in a variety of laboratory environments. Appropriate designs again would be factorial, or, if funding becomes tight, fractional factorial. The next step would be field trials. As McCabe et al. explain, Steve Wunsch moved to Arizona in 1991 with his new electronic market system that competes with the major traditional exchanges in New York and Chicago. Thus opportunities for field experiments seem close at hand.

Among other examples of developmental work, Ferejohn, Forsythe, and Noll (1979) used experiments to examine the characteristics of Station Program Cooperative (a method used by noncommercial television stations in the United States to acquire programming), and to develop alternative bidding procedures. In their preliminary report, they found that the "theoretically superior bidding procedure" was dominated in

important respects by SPC. Plott and Porter (1989) have conducted extensive work on developing market-like institutions for allocation of resources of U.S. National Aeronautics and Space Administration's proposed space station. The future scope for developmental testing seems unbounded.