# 1
# Introduction

## 1.1 Experiments in Economics

Over the last thirty years, there has been a revolutionary change in the methods of economics. For most of the twentieth century, reports of experiments were almost unknown in the literature. Economics—as viewed by economists, and as viewed by professional methodologists—was generally taken to be a nonexperimental science. This understanding of economics is encapsulated in an incidental remark in Milton Friedman's famous essay on the methodology of positive economics—an essay that deeply influenced economists' methodological self-perceptions for at least three decades. Friedman says:

> Unfortunately, we can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences. Generally, we must rely on evidence cast up by the "experiments" that happen to occur.
>
> Friedman (1953, p. 10)

The implication is that the methods of economics, like those of astronomy (in the philosophy of science, the traditional example of a nonexperimental science), are an adaptation to the practical impossibility of controlled experiments. But, from the 1980s onwards, there has been an explosive growth in the use of experimental methods in economics. In terms of most obvious signals, these methods are now accepted as part of the discipline. Experimental research is carried out by many economists around the world. Its results are routinely reported in the major journals. In 2002, Daniel Kahneman and Vernon Smith were awarded the Nobel memorial prize in recognition of their work as pioneers of experimental economics.

Even so, it would be a mistake to think that experimental methods are no longer controversial in economics. Most economists do not conduct experiments and many remain unconvinced of their usefulness, as

experimentalists still often discover when invited to present research papers to general economics audiences. Perhaps more significantly, the apparent consensus that experiments have a legitimate role in economics hides major disagreements about what that role is. Experimental economics is not a unified research program. Indeed, the two Nobel memorial prize winners represent two very different lines of research: Smith is an economist who has developed novel experimental techniques to investigate traditional economic questions about the workings of markets; Kahneman is a psychologist who has used the well-established experimental methods of his discipline to challenge economists' conventional assumptions about the rationality of economic agents. Some commentators have seen these two styles of research as so different that they have reserved the term "experimental economics" for Smith's program, in distinction to the *behavioral economics* of Kahneman's program.[1] We find it more natural to define all forms of experimental research in economics as "experimental economics" and to use the term "behavioral economics" to refer to work, whether experimental or not, that uses psychological hypotheses to explain economic behavior. But whatever terminology one uses, it is undeniable that the research programs pursued by Smith and Kahneman began with different presuppositions and methodologies.

Economists can and do use experimental methods in their own work while rejecting the different methods used by other experimenters. They can and do recognize the value of some programs of experimental research while expressing skepticism about, or even hostility toward, others. There are ongoing disputes about what economics should learn from experimental results, about whether (or in what sense) economic theory can be tested in laboratory experiments, and about how far traditional theory needs to be adapted in the light of experimental results.

Given the speed with which experimental methods have been taken up, and the absence of a tradition of experimental research in economics, the existence of such controversies is hardly surprising. Perhaps for the same reasons, it is not always easy to discern exactly what the disputants are arguing about. In part, these controversies can be seen as normal scientific disagreements about how to interpret new findings. In part, they reflect disagreements about particular features of experimental method. In some cases, however, opponents may be arguing

[1] For example, Loewenstein (1999) criticizes some features of "experimental economics" (by which he means work in Smith's program) "from the vantage-point of behavioral economics."

at cross purposes, failing to appreciate that different types of experiments have different purposes and potentially different methodologies. In other cases, apparent disagreements about experimental method may be the surface indications of much deeper differences between rival understandings of what economics is and how its claims to knowledge are grounded. Because widespread use of experimental methods is so new to the discipline, professional methodologists have only just begun to revise their accounts of the methods of economics to take account of the change. Among the profession generally, there is no recognized set of general methodological principles that can be used to structure these controversies.

This book is the result of our sense that economics needs a methodological assessment of the claims to knowledge that can be derived from the various kinds of experiments that are now being used. Our aim is to offer such an assessment—to describe, appraise, and, where possible, adjudicate between different positions on how experiments do or do not help us to understand the real economic world. In doing so, we hope to enrich the practice and understanding of experimental economics.

We hope to interest at least three kinds of reader: *practicing experimental economists* engaged in these controversies at first hand; *nonexperimental economists* trying to decide how to interpret (or whether to take any notice of) experimental results and the claims that experimentalists make about them; and *philosophers of science* who want to examine the status of the knowledge claims made in economics, or are curious about how a scientific community that once disclaimed experimental methods adapts to their introduction. Clearly, these groups of readers will come to the book with different background knowledge. In the rest of this chapter we provide some basic orientation for our varied readers. In section 1.2, we take a brief look at the history of experiments in economics and ask why economics saw itself for so long as a nonexperimental science. This leads into a discussion of some of the reservations that economists continue to express about experiments, and that feature in ongoing methodological controversies. In section 1.3, we provide outline descriptions of eight experiments, chosen from across the range of experimental economics, broadly interpreted. Our aim here is to give readers who are not familiar with experimental economics a preliminary sense of what this form of research is, and the kinds of claims that its practitioners make. In section 1.4, we use these examples to illustrate the main issues that will be addressed in the rest of the book. Finally, in section 1.5, we explain the stance that we take as authors, as practicing experimental economists writing about the methodology of our own branch of our discipline.

## 1.2 Does Economics Need Experiments?

Perhaps surprisingly, given the general perceptions among economists, the idea that controlled experiments can contribute to economics has a long history.

In an account of the history of experimental economics, Alvin Roth (1995a) uses as his earliest example the work of Daniel and Nicholas Bernoulli on the "St. Petersburg paradox" Bernoulli (1738). The St. Petersburg paradox is a hypothetical problem of decision under risk, in which most people's ideas about reasonable choice contravene the principle of maximizing expected monetary value. In an informal use of experimental methods, Nicholas Bernoulli tried out this decision problem on a famous mathematician to check his own intuitions about it.

We suggest that David Hume is another candidate for the experimental economists' Hall of Fame. Hume's *A Treatise of Human Nature* (1739–40) is now generally regarded as one of the canonical texts of philosophy, but it can also be read as a pioneering work in experimental psychology and decision and game theory. Significantly, the subtitle of Hume's book is: *Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects.* In the preface, Hume describes his work as a study of "the extent and force of human understanding, . . . the nature of the ideas we employ, and of the operations we perform in our reasonings." He undertakes to use the methodology of the natural sciences, investigating the workings of the human mind by "careful and exact experiments, and the observation of those particular effects, which result from its different circumstances and situations" (Hume 1739–40 pp. xv–xvii).[2] In the course of the book, he describes the designs of a series of psychological experiments, and invites his readers to try these out on themselves. Among the results he finds are phenomena that were rediscovered (as so-called anomalies of decision-making behavior) by experimental psychologists and experimental economists in the late twentieth century.[3]

It is particularly significant that neoclassical economics—the orthodox approach to the subject for most of the twentieth century—was, in the first years of its existence, based on experimental research. The pioneers of neoclassical economics were strongly influenced by what were then recent findings of experimental psychology. In launching the "marginal revolution" in economic theory, Stanley Jevons (1871) and Francis Edgeworth (1881) based their analyses of diminishing marginal

[2] Page numbers are from the 1978 edition.

[3] This interpretation of Hume is defended by Sugden (1986, 2006).

utility on psychological findings about the relationship between stimuli and sensations. These authors were well aware of the work of psychophysicists such as Gustav Fechner and Wilhelm Wundt, which they saw as providing the scientific underpinning for the theory of demand. It was only from the beginning of the twentieth century that neoclassical economics separated itself off from experimental psychology, in a self-conscious process initiated by Vilfredo Pareto (1906).[4] Intriguingly, Jevons (1870) may have been the first person to report the results of a controlled economic experiment in a scientific journal. This report, in the second volume of *Nature*, is of a series of experiments carried out by Jevons himself, investigating the relationship between fatigue and the effectiveness of human muscular effort. This was a matter of real economic importance at a time when major civil engineering works were being constructed by men with spades and wheelbarrows. Jevons (1871, pp. 213–16)[5] tells us that he ran these experiments to illustrate "the mode in which some of the laws forming the physical basis of economics might be ascertained." As one might expect of a pioneer of neoclassical economics, Jevons was interested in such maximization problems as determining the optimal size of spade for shifting different materials, and the optimal rate of marching for an army.[6]

Nevertheless, for much of the twentieth century, experimentation was a marginal activity in economics, barely impinging on the consciousness of most economists. With hindsight, it is possible to pick out landmark contributions to experimental economics, some even published in major economics journals; but it is striking that, for many years, very little was done to build on these isolated pieces of work. It seems that they were seen as having curiosity value, rather than as being part of the real business of economics.

For example, an experiment by Louis Thurstone (1931) is now seen as a classic. Thurstone, who was based at the University of Chicago, was one of the leading psychophysicists of his time. Through conversations with his colleague Henry Schultz, an economist doing pathbreaking work on the statistical estimation of demand functions, Thurstone had become aware that the concept of an indifference curve in economic theory had no direct empirical grounding. His experiment attempted to elicit individuals' indifference curves from responses to binary choice problems.

---

[4] For more on this episode in the history of economics, see Maas (2005) and Bruni and Sugden (2007).

[5] Page numbers are from the 1970 edition.

[6] This early exercise in experimental economics was pointed out to us by Harro Maas. The historical and methodological significance of these experiments is discussed in Maas (2005).

Over the following three decades, the project of investigating whether the preferences postulated in theory can be elicited from actual choice behavior was pursued by only a tiny number of economists and decision theorists (see, for example, Mosteller and Nogee 1951; Allais 1953; Davidson et al. 1957; Davidson and Marschak 1959). Maurice Allais's discovery, in the early 1950s, of a systematic divergence between theory and behavior (that we describe in chapter 2) did not much trouble economists for another twenty years.

Similarly, Edward Chamberlin's (1948) investigation of price-determination in an experimental market would appear on any present-day list of great experiments in economics. Chamberlin was a leading industrial economist, famous for his theory of monopolistic competition. His experiment (described in chapter 4) was motivated by his awareness that price theory, despite its formal sophistication, provided no real explanation of how equilibrium is reached in real markets. His results seemed to confirm his hunch that equilibrium would *not* be reached under conditions typical of real-world markets. His paper was published in the *Journal of Political Economy*, but little further work was done for more than a decade. Systematic research on experimental markets was getting under way from the end of the 1950s (see, for example, Sauermann and Selten 1959; Siegel and Fouraker 1960; Smith 1962), but it remained very much a minority taste.[7] It seems that most economists did not think that price theory was in need of experimental support.

Notwithstanding the existence of a few studies now seen as landmarks, it is probable that the large majority of economists saw their subject as fundamentally nonexperimental at least until the last two decades of the twentieth century. For many trained in the third quarter of the century, Friedman's 1953 essay would be their sole excursion into the subject's methodological literature; and echoes of its incidental remark on experiments could also be found in introductory textbooks of the time. For example, consider the following quotation from the 1979 edition[8] of Richard Lipsey's classic textbook:

> Experimental sciences, such as chemistry and some branches of psychology, have an advantage because it is possible to produce relevant

[7] The psychologist Sidney Siegel (1916–61) played an important part in early experimental investigations both of individual decision making (following what would now be called a behavioral approach) and of oligopolistic markets. Innocenti (2008) appraises Siegel's contribution to experimental economics and the loss caused by his premature death.

[8] By 1979, both Vernon Smith and Daniel Kahneman, later to become Nobel laureates of experimental economics, had already completed some of what is now their most famous work.

> evidence through controlled laboratory experiments. Other sciences, such as astronomy and economics, cannot do this.[9]
>
> Lipsey (1979, p. 8)

Even now, one occasionally finds serious writers who echo Friedman's remark. For example, in the abstract of a paper on the methodology of economics published in a recent issue of *Philosophy of Science*, Marcel Boumans (2003, p. 308) asserts: "In the social sciences we hardly can create laboratory conditions, we only can try to find out which kinds of experiments Nature has carried out." Boumans's paper is an extended discussion of the question of how, given the supposed infeasibility of controlled experiments, economics can discover lawlike relationships within its domain of investigation.

Why did economists accept for so long the idea that their discipline was nonexperimental? It is sometimes suggested that the widespread use of experimental methods in economics has become possible only as a result of developments in information technology. It is certainly true that many experimental designs that are now used routinely would have been simply infeasible a few decades ago. The availability of generic software for economics experiments, such as the widely used z-Tree package designed by Urs Fischbacher (2007), has greatly reduced the investment in skills necessary to run computerized experiments. But, as our historical sketch has illustrated, there was no shortage of feasible and potentially informative experimental designs in the first three quarters of the twentieth century—just very little interest in using them. Even in the 1980s—the decade in which experimental methods began to be accepted in economics—many of the most significant experiments used pencil-and-paper technology. What has to be explained is why economists believed for so long that the information that such experiments would produce would not be useful.

Recall that Friedman's comment was that social scientists can seldom test particular predictions in controlled experiments. Since controlled experiments with human subjects are clearly possible, it seems that Friedman must be interpreted as saying that *the kinds of experiments that are possible* cannot be used to test *the kinds of predictions that economics makes*. Lipsey's use of the qualifier "relevant" suggests

[9] As an aside, it is interesting to note that Lipsey draws a sharp distinction between economics and psychology that would now seem harder to defend. But the relationship between experimental economics and experimental psychology has been hotly debated; some, such as Hertwig and Ortmann (2001), point to supposed advantages of economists' techniques; others, such as Loewenstein (1999), argue that experimental economics (of a certain kind) has low external validity, compared with experiments closer to traditions in psychology.

a similar view. Such claims should be understood in relation to two features of mid-twentieth-century economics. First, the domain in which economics was expected to make predictions was, by modern standards, narrow. As is suggested by the then-common use of the term "price theory" as a synonym for "microeconomics," the main focus of microeconomics was on explaining and predicting the values of statistics of aggregate market behavior—in particular, prices and total quantities traded. Macroeconomics worked at an even higher level of aggregation. Thus, the *useful* predictions of economics operated at a level at which, it was thought, direct experimental tests would be enormously costly and perhaps even unethical. The second feature was a prevailing conviction—a conviction for which Friedman (1953) argued strongly—that the "assumptions" of a theory are not claims about how the world is, but merely "as-if" propositions that happen to be useful in deriving predictions. Although price theory was derived from apparently restrictive assumptions about individuals' preferences, those assumptions were not to be interpreted as empirical hypotheses to which the theory was committed. Thus, experiments that purported to "test" the assumptions would be pointless.

A further source of resistance to experiments came from skepticism about whether people's behavior in laboratory or classroom experiments is indicative of their behavior in "real" economic environments—or, as experimentalists now more often say, *in the field*. Friedman again provides an example. As a young economist, he was the coauthor (with Allen Wallis) of a paper on Thurstone's indifference-curve experiment. Wallis and Friedman argue that this experiment is too "artificial" for its results to be reliably transferable to an "economic situation," claiming that "[f]or a satisfactory experiment it is essential that the subject give actual reactions to actual stimuli" (Wallis and Friedman 1942, pp. 179–80).[10]

Friedman seems to have thought that neoclassical price theory, when applied to the "economic situations" for which it was intended, would generally yield successful predictions. But it is surprisingly common for economists to claim that the core theories of their discipline are useful despite being *disconfirmed* by the evidence. This maneuver can be seen in the common idea that theories based on idealized assumptions—for example, the theory of perfect competition, or classical game theory with its assumption of unlimited rationality—provide "benchmarks" for

[10] Viewed from the perspective of modern experimental economics, this involves a non sequitur. A key step in the development of experimental economics has been acceptance of the view, promoted for example by Smith (1982a), that subjects can face and respond to actual economic stimuli even in artificial situations.

understanding the real world. The idea is that we can organize our knowledge of the real world by cataloging its "imperfections" relative to the theory. If one sees a theory in this light, the whole idea of testing it may seem misplaced.

Although probably few economists today would openly dismiss experimental methods out of hand, these (and other) arguments against the validity or usefulness of experiments continue to have resonance in the discipline. Indeed, they are often expressed by experimenters themselves, particularly when criticizing other people's research programs. To illustrate how fundamental questions about the appropriateness of experimental methods remain matters of debate in economics, we look at four recent papers written by well-known economists with experience of experimental research.

In the first paper, Ken Binmore (1999) echoes Wallis and Friedman's reservations about the significance of laboratory results. Binmore's criticisms are directed particularly at the experimental program exemplified by Kahneman's work. Characterizing the main thrust of this program as "denying the validity of orthodox economic reasoning," Binmore urges economists not to be "led by the nose" into accepting its conclusions (pp. F16, F19). He accepts that the behavior of individuals in laboratory experiments is often systematically different from that of the rational agents of economic theory, but rejects the conclusion that the theory has thereby been disconfirmed. Economic theory, he argues, can reasonably be expected to apply only under particular conditions (for example, that decision makers have incentives to deliberate and have had opportunities to learn by experience). Binmore argues that these conditions are not satisfied in the experiments he criticizes. Thus, he concludes, to use the results of such experiments as evidence against economic theory is like claiming to refute chemistry by experiments in which reagents are mixed in dirty test tubes (p. F23).

In the second paper, Steven Levitt and John List (2007) offer guidelines for judging whether laboratory results can be extrapolated to behavior in the field. While Binmore's main concern is with whether the laboratory environment satisfies the conditions presupposed by economic theory, Levitt and List frame their inquiry in terms of how far laboratory experiments capture relevant features of the settings in which economic decisions are made in the field. Their particular concern is with experiments that appear to show that economic agents act on "social preferences" (such as preferences for actions that are construed as fair or trustworthy, or that punish people who have been unfair or untrustworthy). While not proposing the wholesale rejection of any particular class of experiments,

Levitt and List identify various ways in which the "artificiality" of the laboratory might produce results that would not transfer to the field. For example, they argue that laboratory subjects are normally conscious of acting under the scrutiny of experimenters and that, as a result, they may be more inclined to follow moral norms than their counterparts in the field. Echoing an argument used by Friedman (1953), Levitt and List point out that, in many of the environments studied by economists, decision makers are not a representative sample of the population. Instead, people *become* decision makers through processes of selection (for example, to continue in business as a stock-market trader, one has to make profits on one's dealings). These processes might systematically eliminate individuals who act on social preferences. Conversely, standard methods of recruiting volunteers to participate in experiments may select individuals who are predisposed to be cooperative or to seek social approval.

Our third example illustrates a different kind of reservation about experiments. Ariel Rubinstein (2001) writes as a "pure theorist" who has returned from a "short detour" into experimental research. Focusing on decision and game theory, he argues that it is "hopeless and, more importantly, pointless to test the predictions of models in economic theory" (p. 618). For Rubinstein, theoretical models do not generate concrete predictions about behavior in any particular situations. Rather, a model represents, in an abstract form, some "consideration" or "type of argument" that decision makers *might* (not *do*) use. The test of the realism of a model is its intuitive appeal: "[O]ur intuition provides the test. If a phenomenon is robust, we intuitively recognize it as such. It strikes a chord upon us. If we are honest with ourselves, we can feel that it is true" (p. 616). Rubinstein says that, before his detour, he believed that theorists could safely rely on their own intuitions, and so experiments were unnecessary. He now acknowledges that there is a role for experiments as a means of testing whether the theorist's intuitions "ring true" or "make sense" for other people, but *not* as a way of testing theoretical predictions. And, by the end of the paper, he is not completely sure even about that: he leaves it as an open question whether experiments are more reliable than the theorist's "gut feelings" (p. 627).

One of Rubinstein's reasons for thinking this an open question is that "the significance of experimental work relies so heavily on our honesty," with the apparently intended implication that this cannot be relied on (Rubinstein 2001, p. 627). More explicitly, he claims that experimental economics fails to respect certain rules of good scientific method. As one example, he asserts that many experimental economists follow the "problematic practice" of selecting a research question only after

"sifting results ex post: namely, after the results have been gathered" (p. 626). This criticism seems to presuppose a principle that some experimentalists may reject: namely, that the function of experiments is only to *test* hypotheses and intuitions, not to *generate* them. (The latter, presumably, is the role of the theorist.) Here, we suggest, a criticism of the scientific standards of experimental work may conceal a much more fundamental disagreement about how knowledge claims can be grounded.

Our final example is a paper with the provocative title "Experimental economics: science or what?" written by Ken Binmore and Avner Shaked (2007). As in Binmore's 1999 paper, criticism is directed at the inferences that behavioral economists have drawn from experimental results. In this case, however, the criticism is directed not at particular types of experimental designs, but at what Binmore and Shaked argue are inflated claims made on behalf of particular theories. The charge is that some experimental economists use "cherry-picking" methods to appraise their favored theories—in particular, not prespecifying a theory's domain of application before testing it, citing as supporting evidence only those tests that the theory passes, and allowing parameters of the theory to take different values when fitted to different experimental data sets. Whatever one makes of Binmore and Shaked's view of particular theories in behavioral economics, their paper draws attention to important and unresolved methodological questions for experimental economics. When experimental evidence reveals systematic deviations from previously received theory, how should economists go about incorporating those findings into new theories? How far can one legitimately generalize from narrowly defined classes of experiment to the whole range of cases to which economic theories are expected to apply?

For the purposes of this introductory chapter, it is sufficient to recognize that the role of experimental methods in economics remains controversial. These controversies provide the context and raison d'être for our book.

## 1.3 The Practice of Experimental Economics

Before commencing a discussion of experimental economics, it is important to have some picture of what it involves. Since we do not presume that all readers will be familiar with the field, we start by illustrating some of the things that experimenters do. Since the literature is now vast, we cannot sensibly attempt a review of experimental economics

in the round.[11] Instead, our strategy is to describe a few published papers that exemplify some of the main genres of experimental research. We aim to illustrate the sorts of questions that have motivated experimenters in economics and the methods they have used to tackle them; we will describe some of the main results reported in these papers and note broader claims made by the authors. In this section, our aim is strictly descriptive; we wish to give a compact account of what various researchers did, found, and wrote, while (for the moment) avoiding any evaluation of it.

Many of the things that experimental economists now do have close parallels with work that has older roots in the traditions of experimental psychology. With this in mind, we begin with two illustrations of research conducted by psychologists. The first of these, due to Amos Tversky and Daniel Kahneman (1981), is a laboratory experiment investigating individual decision making with a particular focus on decisions involving risk.

**Illustration 1 (Tversky and Kahneman (1981), "The framing of decisions and psychology of choice," *Science*).** Tversky and Kahneman present the results of experiments investigating whether small changes in the description of the alternatives available in a decision problem, which apparently leave the logical structure of the decision problem unchanged, might nevertheless affect what is chosen. They find that seemingly inconsequential changes in the "framing" of decision problems can have a substantial impact on choices. Here is one example. They compare the behavior of two groups of subjects. One group of 152 subjects was confronted with the following choice problem:

> Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows.
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
>
> Which of the two programs would you favor?

[11] The eight chapters of Kagel and Roth (1995) provide heroically comprehensive guides to the main areas of experimental economics as they existed up to the mid 1990s. But, as the bibliographies of many of the chapters ran to several pages then and the subsequent literature has grown very rapidly, a similarly comprehensive record of the whole of experimental economics would now require several volumes. For example, Camerer (2003), which runs to over five hundred pages, surveys the experimental literature on games, a topic that occupied just three of the chapters of Kagel and Roth (1995).

A second group of 155 subjects were presented with the identical scenario except that the outcomes of the two alternatives were described as numbers of lives lost (out of 600) instead of lives saved. Hence, the two alternatives presented to the second group were:

> If Program C is adopted 400 people will die.
>
> If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Placed side by side, it is easy to see that the choices offered to the two groups are logically equivalent. Nevertheless, Tversky and Kahneman report very different behavior across the groups: in the first group, the majority (72 %) preferred Program A; while in the second group, the majority (78 %) preferred Program D.

The decision makers facing these tasks were students at either Stanford University or the University of British Columbia and the decisions were presented as questionnaires conducted in classrooms. While the example just described was presented to the students as a purely hypothetical choice, other choices made by subjects in their studies involved the possibility of real monetary payoffs. Tversky and Kahneman (1981, p. 453) interpret their evidence as, on the one hand, demonstrating violations of principles of coherence implicit in rational-choice theory, and, on the other, as providing clues to "psychological principles that govern the perception of decision problems and the evaluation of options."

Our second illustration, due to James Bryan and Mary Ann Test (1967), reports a set of experiments designed to explore whether individuals are more likely to engage in altruistic behavior when they see examples of other people doing so.

**Illustration 2 (Bryan and Test (1967), "Models and helping: naturalistic studies in aiding behavior," *Journal of Personality and Social Psychology*).** This paper reports four experiments. One of them compared two conditions that we will call the "baseline" and the "treatment." In the baseline condition, a young female undergraduate was commissioned to stand by a car (a 1964 Mustang) with a flat tire in a residential area of Los Angeles. The setting was staged so as to make it easy for other passing drivers to see the woman, the flat tire, and the presence of an inflated tire leaning against the car. This baseline treatment was run on two successive Saturday afternoons, across a time interval long enough to allow exactly 1,000 cars to pass on each day, and the experimenters recorded how many drivers stopped to offer help. The treatment condition was exactly the same except that a second car (an Oldsmobile) with a flat tire

and a young woman standing by it was positioned $\frac{1}{4}$ of a mile upstream of the traffic flow. This car, however, was already raised on a jack and a man was changing the tire as the woman watched. The order of the two conditions was reversed across the two Saturdays to "counterbalance" the design. Bryan and Test report that significantly more vehicles stopped in the treatment condition (58 out of 2,000, compared with 35 out of 2,000 in the control condition).

Bryan and Test's other three experiments involved collections for the Salvation Army on busy shopping streets. In one of these experiments, a person wearing a Salvation Army uniform stood with a collection box (or "kettle") outside the main entrance of a large department store in Princeton, New Jersey. The collector could ring a bell to signal their presence, but they were instructed to make no verbal appeals for donations and to make no responses to actual donations. Against this backdrop, the experimenters then simulated altruistic acts in the following way: every sixty seconds, a second person in the employ of the experimenters (dressed as a white-collar worker) would approach the collector from within the store, deposit 5 cents, and then head away. Observations were collected for a total of 365 such minutes. Bryan and Test hypothesized that people passing in the vicinity of the collection point when the experimental stooge made a donation would be more likely to donate themselves. To test this hypothesis, they compared the frequency of donations occurring in the twenty-second "window" immediately following the stooge's donation with those occurring in the subsequent twenty-second window. They recorded a total of sixty-nine donations occurring in the first window and only forty-three in the second. The difference is strongly statistically significant.

Bryan and Test interpret their findings as providing support for the hypothesis that "observation of altruistic activity will increase such behavior among observers" (Bryan and Test 1967, p. 403). A feature of all four of their experiments is that the participants whose decisions were being observed were simply passers-by who, it was intended, should have no knowledge of their participation in an experiment at the point they took their decision to help (or not).

Our third illustration also investigates individual decision making, but this case provides an example of a more recent study, designed and run by economists. This experiment gathers data on individual risk preferences using two quite standard approaches: individuals are asked to make straight choices between pairs of gambles; they also state selling prices for various gambles. Data of this sort has been gathered for a number of purposes. In some cases, interest centers on measurement

of preference parameters in particular theories; in other cases, preference data are gathered with a view to testing particular preference theories.[12] In our third illustration, however, the primary purpose was to test the reliability of what had been a widely used experimental incentive mechanism: the *binary lottery procedure*.

The binary lottery procedure works in the following way. Suppose subjects taking part in an experiment earn points as a consequence of completing specific tasks: those tasks might, for example, be participation in strategic games, and the points would be the payoffs resulting from the strategy combinations of the players. One way to incentivize such tasks is to translate the points won by a subject into a money payment at some exchange rate. However, the binary lottery procedure instead translates points into a probability of winning a known prize in a binary lottery (i.e., one in which the subject either wins or receives nothing). The attraction of this approach is that if individuals' preferences satisfy certain conditions, subjects in an experiment implementing this procedure should make risk-neutral decisions. The ability to control preferences, by screening out attitudes to risk, could be very useful, if it works. An experiment reported by Reinhard Selten et al. (1999) was designed to test whether it does. It provides an example of how experiments can be used to investigate the properties of devices used by other experiments. This genre has a counterpart in the natural sciences, where experiments are sometimes used, for example, to test the reliability of measuring instruments.

**Illustration 3 (Selten et al. (1999), "Money does not induce risk neutral behavior, but binary lotteries do even worse," *Theory and Decision*).** In this experiment, each subject made a series of thirty-six pairwise choices between two lotteries. They also stated minimum selling prices for fourteen lotteries and the experimenter then generated a random offer to determine whether the subject kept or sold this lottery. Subjects were randomly allocated across two treatments, one of which involved a binary lottery procedure. In both treatments, after every pair of tasks—called a "round"—subjects received a payoff in points determined by their decisions and the resolution of lotteries from that round. In the binary lottery group the point payoff from the round then determined a probability used to play a "grand" lottery for a fixed prize. Subjects in the other condition simply had their point payoff from each round converted to a money payoff using an exchange rate designed to

[12] Holt and Laury (2002) is an example of the first case and Camerer (1989) of the second. Camerer (1995) and Starmer (2000) provide surveys.

equalize expected earnings across the groups. A subset of subjects facing each of these conditions also had access to computerized calculators which would report, on request, measures including a gamble's expected value.

A total of 144 subjects, recruited from the student population at the University of Bonn, took part. Sessions lasted about ninety minutes and subjects earned (on average) between DM15.50 and DM39.50 in this time. To analyze the data, the authors constructed a measure of the extent to which each individual's decisions depart from expected value maximization. Their primary conclusion was that subjects in the treatment using the binary lottery procedure departed from expected value maximization significantly more than those subjects who received direct money payments. Hence they concluded that "the use of payoffs in binary lottery tickets in experiments is counterproductive" (Selten et al. 1999, p. 225).

Another major line of enquiry in experimental economics focuses on strategic behavior in stylized games. We have three illustrations of work in this genre. The first, due to Jacob Goeree and Charles Holt (2001), reports behavior across a range of games and concludes that predictions based on Nash equilibrium work well in some cases but not in others.

**Illustration 4 (Goeree and Holt (2001), "Ten little treasures of game theory, and ten intuitive contradictions," *American Economic Review*).** This paper examines behavior in twenty two-player games, consisting of ten matched pairs of games: in each such pair, there is one game where observed play corresponds well with Nash equilibrium predictions (these are the ten "treasures" of the paper's title) and another game for which the researchers predict (and find) that behavior deviates markedly from such predictions (these are the "intuitive contradictions").

Participants were undergraduate students recruited from economics classes at the University of Virginia. Subjects took part in sessions involving ten people. Prior to making decisions in the games reported in the paper, the subjects played a repeated two-person game with random matching. This was intended partly to familiarize participants with general features of the task environment, but in a different game from the subsequent one-shot tasks. After this, subjects responded to a set of one-shot games: these were a subset of the pairs of treasure/contradiction treatments reported in the paper as a whole. Sessions lasted two hours and subjects were paid an average of $35 (including a flat fee of $6 for turning up).

Here is one example from the study based on the "traveler's dilemma" game due to Basu (1994). In this game, two players simultaneously select an integer in the range 180–300. The payoffs are then determined in the

following way: both players receive the lower of the two values, but, in addition, a "transfer" amount $T > 0$ is added to the payoff of the player who set the lower amount, and subtracted from the payoff of the player who set the higher amount. With $T > 0$, a player would maximize their own payoff if they managed to state a number just one less than their opponent. So if I am playing the game and I expect the other player to play 300, I should play 299. But if they expect me to play 299, then they should play 298. As Goeree and Holt explain, this type of reasoning implies that the game has a unique Nash equilibrium, in which both players select the number 180. While this equilibrium prediction does not depend on the value of $T$ (provided $T > 0$), the authors suggest that behavior might depend on $T$ because, since $T$ is the cost of being underbid, Nash equilibrium predictions might work better when this cost is relatively high. They test this conjecture by observing play across two implementations of the game with $T$ set either High ($T = 180$) or Low ($T = 5$). They report data based on the decisions of fifty subjects who were paired to make decisions in both the "High-$T$" and "Low-$T$" versions (the two games were presented in random order and separated by having subjects make decisions for other pairs of treasure/contradiction games reported in the paper). The authors report very high conformity with Nash predictions in the High-$T$ condition (around 80% of subjects state 180) and very low conformity with Nash in the Low-$T$ condition. In the latter case, only around 20% play the Nash strategy; moreover, relative to the Nash prediction, the majority of subjects chose values at the opposite end of the strategy space.

Our next example is an investigation of contributions to public goods. Like many experiments on this issue, the design is built around a device known as the voluntary-contributions mechanism. In an experiment that uses this mechanism, each subject in a group has to divide an endowment between a private good and a public good. Payments to each subject are determined by the decisions of all group members in such a way that, for a subject who seeks to maximize their own monetary payoff, it is a dominant strategy to allocate their whole endowment to the private good, even though all group members would be better off if all contributed their whole endowment to the public good (see box 2.3 (p. 58) for more detail). Typical findings from the many studies that have used this device are that, in a finitely repeated game, many subjects begin by contributing around half of their endowment to the public good, but that average contributions decline toward zero with repetition. The study by Ernst Fehr and Simon Gächter (2000) is set against this backdrop. It

investigates whether the opportunity for players to punish each other affects contributions in a public-goods game.

**Illustration 5 (Fehr and Gächter (2000), "Cooperation and punishment in public goods experiments," *American Economic Review*).** In this experiment, subjects played repeated rounds in groups of four, selected from a larger pool of participants. Some groups played under a "strangers" protocol, meaning that groups were selected at random from the larger pool separately for each round; others played under a "partners" protocol meaning that, although selected at random initially, the groupings stayed the same across rounds. (See box 2.8 (p. 88) and section 2.7 for further discussion of these protocols.) Subjects were told which of these conditions applied but, under both conditions, they interacted anonymously via computer terminals. Subjects knew that they would receive a money payoff at the end of the experiment determined by their final holdings of "tokens" from all rounds.

The main innovation of Fehr and Gächter's design was to introduce the possibility of punishment. Subjects in both protocols played ten rounds without punishment opportunities and ten rounds with them (the order of these conditions was varied, to control for order effects). When punishment opportunities were not available, in a given round, subjects played a voluntary-contributions game of the kind described above. But, when punishment opportunities were available, the voluntary-contributions stage was followed by a further stage in which, after being informed of contributions made to the public good by their fellow group members, subjects could award "punishment points" to them. Each punishment point reduced the payoff of the punished subject for that round by 10%, but punishing was also costly to subjects awarding it. Because of the latter feature, a standard game-theoretic argument implies that, if subjects care only about their own money payoffs, punishment points will never be assigned and, therefore, the opportunity to assign them will have no effect on contributions to the public good.

The no-punishment conditions replicated well-known findings of experiments using the voluntary-contributions mechanism: significant contributions were observed in early periods, but these decayed across rounds and approached full free-riding by round ten. This serves as a baseline against which to compare behavior in the punishment conditions. Although punishing was costly, punishing was observed, even in the strangers condition. The opportunity to punish had a significant impact on contributions behavior, with differences between strangers and partners conditions: in the strangers treatment with punishment, contributions no longer converged toward free-riding, but there was

considerable variation in behavior and "no stable behavioral regularity regarding individual contributions" (Fehr and Gächter 2000, p. 986) emerged. In the partners condition with punishment, behavior appeared to converge toward the Pareto-efficient (full contribution) outcome.

Our third "games" example, a study by John Morgan et al. (2006), straddles the literature on games and markets. The experiment tests the predictions of a particular game-theoretic model: a "clearinghouse" model of pricing behavior among oligopolistic firms.

**Illustration 6 (Morgan et al. (2006), "An experimental study of price dispersion," *Games and Economic Behavior*).** The paper begins by setting out a simple clearinghouse model; a variant of that due to Varian (1980). In this model, $n$ identical firms (facing constant marginal costs) engage in price competition to sell a homogenous product. On the demand side there are two types of consumers: a fraction $\lambda$ of "informed" consumers are assumed to buy from the firm setting the lowest price; the remaining consumers are "loyal" in the sense that their demand is equally distributed among all firms who offer prices below a critical reservation value. Morgan et al. show that when firms are risk-neutral profit maximizers, the game has a unique symmetric mixed strategy equilibrium. They also highlight two comparative static implications of the model that are key to their tests: increasing $\lambda$ is predicted to reduce the prices faced by both types of consumers; whereas increasing $n$ is predicted to impact differentially for the two types, *reducing* prices for informed consumers and *increasing* those for loyal consumers.

The experiment was run as a series of sessions, each involving subjects from the student population at the University of Nottingham. In each session, subjects sat at computer terminals and participated in experimental markets structured to mimic various features of the theoretical model. The participants played the role of sellers in these markets and their task was to set their firm's price in each of ninety market periods. The number of sellers in each market was held constant within a session, but varied between sessions (half the sessions had two sellers per market, the rest had four). At the start of each period, sellers were randomly grouped into markets (of either two or four sellers depending on session). They then simultaneously set their "prices" by selecting a number ranging from 0 to 100. The demand side of the market was simulated using computerized consumers. Sellers knew that in every period there were six consumers who would buy twelve units each; a known fraction of these consumers, being "informed" in the sense of the model, would buy from the seller with the lowest price, while the demand of the others was allocated evenly across sellers regardless of their prices. At the end

of each period, each player learned the prices set by competing sellers and the resulting pattern of sales. They earned a number of points equal to their own sales level times their price. The ninety periods were divided into three phases of thirty periods and the number of informed buyers was varied across phases: for the first and last phase, half of the buyers were informed; for the intermediate phase, five out of six were informed. Sessions lasted about ninety minutes and subjects were paid a flat fee for attending plus a money reward that increased by a penny for every 100 points earned in the experiment.

The authors report mixed success for the predictions of the theoretical model. On the negative side, there are significant discrepancies between the predicted and observed price distributions: relative to theoretical predictions, observed prices are too high in two-seller treatments and too dispersed in four-seller treatments. In contrast, the comparative static predictions relating to changes in both the number of sellers and the proportion of informed consumers are broadly supported in the data.

In our next illustration, due to Smith et al. (1988), the researchers created and observed trading in an experimental asset market, with a view to studying the incidence of speculative bubbles. In economics, there has been much debate about the extent to which volatility in financial markets reflects speculation-fueled bubbles, i.e., deviations from "fundamental" asset values. An obstacle to reaching clear conclusions in these debates through the analysis of field data is the fact that some fundamentals are typically unobserved in such data. The following study came at the problem by creating a market setting in which fundamentals were controlled (and so known) by the experimenter. Consequently, the experimenters hoped to observe the extent and persistence of any deviations between prices and fundamental values.

**Illustration 7 (Smith et al. (1988), "Bubbles, crashes and endogenous expectations in experimental spot asset markets," *Econometrica*).** This paper reports twenty-seven experiments investigating experimental markets in which participants had the opportunity to trade "assets." While many details vary across the different experiments, several structural features are common across them and for the most part we focus on those.

At the beginning of an experiment, participants were each endowed with experimental "assets" and "cash." They then took part in a series of (usually fifteen) market periods in which they had opportunities to buy assets for cash or to sell them and receive cash in return. The markets were organized as computerized *double auctions*. At any moment during a market period, individual participants could submit offers to buy a

unit (by stating a price they were willing to pay) or to sell a unit (by stating a price they were willing to accept). All bids were subject to an improvement rule: any new bid to buy (sell) must be higher (lower) than the current best offer on the market. An asset would be traded when one participant accepted a standing offer (to buy or to sell) posted by another trader. At the end of each market period, every asset generated a dividend payoff to its current owner, which was added to their cash balance. The value of this dividend was determined randomly from a known probability distribution over dividends, the parameters of which were fixed across rounds. At the end of the experiment, subjects received a real dollar payment based on their final cash balance (incorporating all payments and income from trades, plus cumulative dividend payments.)

Earlier research had investigated some aspects of asset trading in experimental markets, but the setup of this study featured two main departures from previous designs. First, the assets were "long-lived" in the sense of generating a stream of dividends across multiple market periods; second, the expected dividend was held constant across traders. The second feature meant that if traders were risk neutral (or had a common risk attitude), and acted on the basis of rational expectations, then there would be no incentives for any asset trading to occur. That the design allowed trade across an extended number of periods was intended to permit examination of the dynamics of trade volumes and prices, should there be significant levels of trade contrary to the theoretical prediction. Variations across the experiments explored various issues, including the extent to which the "experience" level of the traders affected the conformity of market behavior with predictions.

Smith et al. report not just that trade occurred but also that it was common for their markets to exhibit "bubbles": that is, sustained periods during which assets traded at prices significantly above their expected returns. Such bubbles were typically followed by crashes, with prices and volumes of trade collapsing, near the final period. They also report that trader experience had some tendency to attenuate bubbling phenomena.

Our final illustration, due to Sheryl Ball et al. (2001), also concerns an experimental market. In this case, the interest is in examining whether the status of market participants influences the distribution of surplus in a market constructed to have multiple price-equilibria.

**Illustration 8 (Ball et al. (2001), "Status in markets," *Quarterly Journal of Economics*).** This paper reports the results from some experimental markets. Each market involved between ten and sixteen participants divided randomly between the roles of buyer and seller. In each market period, each seller was endowed with two units of an experimental

good, they were each given a "private cost" and could make money in the experiment by selling their units at prices above their own costs. Individual buyers were each given a reservation value and knew that they could make money by buying units at prices below their reserve. Each subject knew only their own cost or reserve, but in fact in a given market the cost was the same for all sellers and the reserve was the same for all buyers. Since the cost was set below the reserve, any buyer–seller pair could, in principle, undertake a mutually profitable trade at any price in the cost–reserve interval. Markets were run for eleven periods, each by an auctioneer who alternated between inviting a randomly selected buyer to bid and inviting a randomly selected seller to state an ask. The first period was a practice but the returns to trade in the remaining ten periods contributed toward the subject's final payoff (subjects earned an average of around $17 including a $5 turn-up fee).

The primary objective of the research was to test the hypothesis that "In markets where sellers have higher status, the distribution of equilibrium prices will be higher than in markets where buyers have higher status" (p. 165). To this end, the researchers compared matched pairs of markets, identical in terms of costs and reserves, in which participants had been exposed to a prior manipulation intended to induce differential statuses. Two procedures were used for this. One, the "awarded" status procedure, required subjects to participate in a quiz with obscure answers. The experimenters awarded gold stars to half of the subjects (the "high-status" group) while the other (the "low-status" group) were required to observe and applaud. It was intended that, from the subjects' point of view, it should appear that those getting the stars had earned them through success in the quiz, though in fact stars were assigned randomly (via cover of an opaque scoring system for the quiz). Consequently, awarded status should not have been correlated with personal characteristics such as knowledge or intelligence. The second method for determining status, the "random" status procedure, similarly selected half of the subjects for the public award of stars, but in this case using a procedure intended to make the randomness of the status assignment transparent to subjects.

Ball et al. report behavior in markets under four conditions: in roughly half of the markets, the buyers (respectively sellers) were the high-status group; and for each of these conditions, there were matched pairs of markets which varied according to whether status had been determined by the awarded or random procedures. They report that, in the aggregate, mean earnings tended to be significantly higher for the high-status side of the market. Moreover, the effect appeared to operate even when status

was awarded in a transparently random fashion. In the conclusion to the paper, the authors comment that:

> Our results show that in a competitive market environment, status can have an effect on price and the allocation of resources. That a status treatment that is so obviously superficial could have such an effect on behavior strengthens our belief that status plays an important role in real-world interactions.
>
> Ball et al. (2000, p. 181)

## 1.4 The Illustrations and the Structure of the Book

The experiments presented in section 1.3 illustrate many of the activities that experimental economists undertake and allow us to introduce the main issues that we discuss in later chapters. A central theme of the book will be a distinction between two ways of viewing experiments: as providing tests of theories and as investigating empirical regularities. As we will explain, these categories are not mutually exclusive, nor are they exhaustive; but, nevertheless, they play a useful organizing role.

In chapters 2 and 3, we consider the classic role of experiments in science: namely, *testing theories*. Among our examples, Morgan et al. (2006) and Goeree and Holt (2001) most clearly present the main objective of their study as being to test a theory. In several other cases, theory testing is one way of reading the results, though, as we will discuss later, it is not the only way. For example, Tversky and Kahneman's (1981) investigation of framing effects can be seen as a test of the principle, embedded in consequentialist theories of choice, that logically equivalent redescription of a decision problem will not affect behavior; Fehr and Gächter's (2000) study can be seen as a test of a game-theoretic prediction that subjects will neither contribute nor punish in their setup; and Smith et al.'s (1988) study can be seen as testing a theory that predicts the absence of trade in their laboratory asset markets.

Notwithstanding the prominent place of theory testing in accounts of scientific method, the theory-testing function of economics experiments is not straightforward. Perhaps the most basic reason for this is that the relationship between economic theorizing and empirical claims is itself indirect. Most of the explicit activity of economic theorists is deductive and involves the manipulation or analysis of formal models and definitions. One way in which theorists present their research is by stating assumptions and deriving conclusions from them, using mathematical or logical arguments. Conclusions are then presented as theorems, with

the form of conditional statements to the effect that *if* the assumptions held in the world, *then* the conclusions would also hold. However, it is perhaps now more common for theorists to proceed in a different way, by postulating or building a model world populated by entities defined in formal terms and then deriving theorems that make *unconditional* statements about the *model world*. There are differences between these two ways of presenting theory but the important thing about them, for present purposes, is something that they share. For both presentations, if one sees the assertions made by the theory as consisting *only* of the theorems, then there is no scope for testing the theory, where this is conceived as an activity that involves empirical observation. For the first presentation, this is because the theorems only assert that the conclusions hold if the assumptions do. That the conclusions follow from the assumptions is (provided the theorist has not slipped up in the formal argument) a matter of logic. For the second presentation, it is because the theorems state properties of the model world, not the actual world.

However, we take it that most of our readers and most theorists would, like us, not be content with a view of economic theory that immunizes it not just from experimental testing but also from *all* empirical testing. Our starting point is, therefore, a view of economic theory that does not see its assertions as limited only to theorems whose truth can be established by nonempirical methods. Even though deductions made within particular formal models are not themselves subject to empirical test, if the models are to assist economics as a science, there must be some empirical claims associated with them. *These* claims are ones with which an enterprise of theory testing can sensibly be concerned; they stem not from formal theorems alone but from applications of the models whose properties those theorems establish.

For example, the study by Morgan et al. (2006) concerns a particular game-theoretic model. In that model, the game can be described in formal terms, using an abstract set of players, abstract sets of strategies for the players, and payoff functions defined on the Cartesian product of the strategy sets. Given this formal description, it is a theorem that the game has a unique Nash equilibrium and that this equilibrium has a certain form, consisting of particular mixed (i.e., random) strategies. An application of the model, sufficient to render testable claims, must associate the "players" of the model with some real agents and the "strategies" with certain options open to those agents; and it must endorse as a prediction some solution concept for the game. If this solution concept is the mixed strategy Nash equilibrium, then it is necessary to specify what observations would be taken as conforming, or not conforming, to mixed strategy equilibrium play. The application implicitly made by

Morgan et al. associates players with subjects, strategies with numbers representing prices, and takes the distribution across players of "prices" chosen as indicative or otherwise of mixed strategy Nash equilibrium play.[13]

The need for a theory to be applicable before it can be tested gives rise to many of the questions we discuss in the book, starting with those in chapters 2 and 3. Put loosely, the questions considered by these chapters are, respectively, *where* and *how* the theory should be applied.

Chapter 2 considers whether laboratories provide appropriate testing grounds for economic theories; and, if so, whether certain design features are mandated by particular views about the theory. Each of these questions can be motivated with reference to a particular aspect of the instrumentalist methodology espoused by Friedman (1953). Friedman contended that a theory should be judged by the success of its predictions *within the domain in which they are intended to apply*. For Friedman, it does not matter if, say, the theory of the profit-maximizing firm does not fit well with discussions in firms in which managers take more note of average costs than they do of marginal costs, as long as the theory predicts accurately how market prices respond to changing conditions. It is the latter, not the content of boardroom discussion, which is the intended domain of the theory. An observer sympathetic to this view might question the testing of economic theories in the laboratory. For example, Morgan et al. (2006) purport to test a theory of price dispersion, presumably intended to apply to markets made up of real firms and their customers. The observer might ask, Is it legitimate to test the theory in a laboratory environment with students playing the role of firms and the demand side of the market simulated by computer?

The questions addressed by chapter 2 are similar to this one, though couched in a more general form. The first issue is whether the domain of the theory excludes the laboratory, as suggested by the traditional mid-twentieth-century view of economics as nonexperimental. The second is whether, even if economic theory can legitimately be tested in the laboratory, as is now widely accepted, the nature of its domain requires particular types of experimental designs to be used. To motivate this question, consider again Binmore's (1999) discussion of the research of Kahneman and Tversky. Binmore argues that economics experiments

[13] An alternative application of mixed strategy Nash equilibria would specify that, over repeat play, *each* player's observed behavior is consistent with *that* player randomizing over their pure strategies, using the probabilities specified by the equilibrium mixed strategies. Morgan et al. (2006, pp. 150–51) consider this, but the majority of their analysis focuses on conformity of the distribution of play *across* players with the equilibrium mixed strategies.

ought to conform to certain design principles involving task simplicity, incentives, and learning opportunities. In support of this, he writes that "there is no point in testing economic propositions in circumstances to which they should not reasonably be expected to apply" (Binmore 1999, p. F23), a sentiment which, to the extent of restricting tests of economic theory to a particular domain, is reminiscent of Friedman, even if casting such a sentiment as a restriction on appropriate experimental designs is different. A related question to that of whether certain design features are required for tests of the theory is what one should conclude about the theory if its performance in the laboratory depends in systematic ways on such features. This question has been considered by another prominent experimental economist, Charles Plott, in formulating the *discovered preference hypothesis* (Plott 1996). Plott's view suggests similar boundaries for the domain of the theory to those envisaged by Binmore. Chapter 2 offers a framework for the assessment of such positions.

Chapter 3 turns to the implications for experimental economics of one of the fundamental problems of empirical science: the *Duhem–Quine problem* (Duhem 1906; Quine 1951, 1953). The problem is that, since theories must be applied before they can be tested, single theoretical hypotheses can never be tested in isolation. Supplementary assumptions are always involved in bringing a particular theoretical hypothesis into confrontation with data. As a result, if the data seem unfavorable to the hypothesis, it is never completely clear whether the hypothesis itself is at fault or whether some of the supplementary assumptions were invalid. For example, Smith et al.'s (1988) experiment can be seen as testing a hypothesis, derived from the theory of asset markets, that agents who have the same endowments of, information about, and preferences over particular assets will not trade. The observation that trade occurred in its asset markets certainly tells against the proposition that there would be no trade in them. But does it falsify the theoretical hypothesis? Or might subjects have varied in their reasons for holding the assets, perhaps because of differing attitudes to risk or differing beliefs about the likely trading behavior of other subjects? Or did some subjects just trade for fun? Or vary in their views about the truthfulness of the experimenter's reports of the assets' returns? Or, as a result of integrating experimental endowments with nonlaboratory wealth, differ in their endowments? To raise these questions is not to criticize Smith et al.; it is simply to illustrate how, even in a well-controlled experiment, there is more than one conceivable interpretation of given observations.

Chapter 3 discusses the implications of this fact for the view of experiments as tests of theories, especially where those theories are core principles deeply embedded in economic models. As an example of the latter

kind, consider the treatments with punishment opportunities in Fehr and Gächter's (2000) experiment. If individual subjects want only to accumulate money for themselves, then, in a game-theoretic model of the setup created by these treatments, the unique subgame-perfect equilibrium precludes both punishment and contribution to the public good. If, as was the case, contribution and punishment are observed, does this tell against the game-theoretic concept of subgame perfection, against a particular assumption about players' objectives, or against some other assumption involved in applying a game-theoretic model to a specific situation? As such questions are unlikely to be answered by a single experiment, we require principles for the conduct and evaluation of research programs. Chapter 3 discusses precisely this form of response to the Duhem–Quine problem.

Chapter 4 broadens the perspective of the book by introducing a view of experiments that allows them to be something other than theory-testing devices. As noted earlier, an alternative interpretation sees experiments as contributing to the investigation of *empirical regularities*. Such investigation may involve attempts to "sharpen" the regularities, as well as to contribute to a search for explanations of them. Among our examples, the studies that fit this reading most immediately may be Bryan and Test (1967) and Ball et al. (2001). These studies concern supposed relationships between social experience and charitable behavior and between status and market outcomes, respectively. Several of the studies that we discussed in the role of theory tests can also be seen as investigations of empirical regularities (and, in some cases, it is this reading that the authors stress). For example, Fehr and Gächter (2000) can be read as an investigation of the part punishment opportunities play in supporting social norms of cooperation; Tversky and Kahneman (1981) as an investigation of the effects of descriptions on perceptions; and Smith et al. (1988) as studying the determinants of asset price bubbles. Subsequent research motivated by these studies has followed both readings. For example, Fehr and Gächter (2000), together with many other reports of experiments involving public-goods problems and other games, has spawned an extensive theoretical literature that develops models of social preferences.[14] But there has also been a major program of experimental research (surveyed in Camerer (2003, chapter 2)) that investigates the robustness of the behavioral regularities to changes in cultural context, design features, and other factors. Some of this research can also be seen as a follow-up to Tversky and Kahneman (1981), since

[14] For surveys, see Bolton (1998), Fehr and Fischbacher (2002), Fehr and Schmidt (2003), Camerer (2003, section 2.8), Bardsley and Sugden (2006), and Sugden (forthcoming).

it considers the effects of task framing on behavior in collective choice problems.

The existence of sustained programs of experimental research into particular regularities in the behavior of subjects brings out an important point about experiments, when read as investigations of empirical regularities. They are not blind searches, conducted in the hope of stumbling on regularities. For example, Smith et al. could reasonably have conjectured that bubbles might form in their laboratory asset markets because of the widespread perception that they do form in nonlaboratory markets in which shares and real estate are traded; and Bryan and Test could have conjectured, for example from the use of auctions as fund-raising devices by charities, that positive examples stimulate contributions. But such conjectures are notoriously difficult to confirm in field research, because there are so many potentially confounding factors.[15] Hence the potential role for experiments to refine the understanding of what regularities in behavior there are. In view of our claim that experiments, read as contributions to investigation of empirical regularities, are not (or at least should not be) blind searches, we prefer to think of this reading of experiments as postulating a role of *regularity refinement* or *regularity confirmation* rather than haphazard regularity hunting.[16]

When experiments are used in this way, they can take on roles that would otherwise be played by theoretical models. For example, suppose an economist is asked how prices and quantities traded in some market would be affected by some exogenous change in circumstances. The traditional response would be to build a theoretical model of the market, using components from some received economic theory (such as the Marshallian or Walrasian theory of market equilibrium), and to manipulate the model in ways that correspond with the relevant exogenous changes. But if it is known that certain types of experimental market designs reliably generate patterns of behavior that are similar to those observed in real-world markets, another strategy of investigation is possible. The real-world question that the economist is trying to answer can be represented not by a manipulation of a theoretical model but by alternative treatments in an experimental design; the answer can be arrived at not by interpreting a mathematical theorem but by interpreting an experimental finding. To put this in another way, an experiment is being

[15] One example is the debate about whether famous early "bubbles" really were bubbles. For example, compare Kindleberger (1996) with Garber (2000). See also the "Symposium on Bubbles" in the Spring 1990 issue of *Journal of Economic Perspectives*.

[16] Roth (1995a, p. 22) discusses "searching for facts" as a role for experiments and, like us, stresses the systematic nature of the activity, when fruitfully conducted. He sees part of its value as being to make possible the formulation of theories.

used *as a model*. The work of Ball et al. (2001) can be thought of in this way. The research question, one might say, is whether status differences affect the terms of trade in markets. Ball et al. try to answer this question by setting up an experimental market, using components that are standard in experimental research, but adding new design features that are intended to model status.

One important application of this research strategy uses experiments as "test beds" or "wind tunnels" for investigating the likely properties of new market institutions, prior to their being introduced for real.[17] The starting point for this strategy is the claim that, when certain general design principles are followed, behavior in experimental markets tends to be similar to behavior in their real-world counterparts. If this claim is supported by experience, the performance of an experimental model of a new market institution can reasonably be treated as informative about the likely performance of the institution itself. This test-bed strategy will be discussed in chapter 4.

More generally, chapter 4 discusses experiments as contributors to *inductive* reasoning in economics. One issue that immediately arises is the relationship between the theory testing and regularity-refining reading of experiments outlined above. Is it, for example, coherent to suggest, as we did earlier, that the same experiments can be read either way? Are the rules of theory testing and regularity-refining different? Indeed, what are appropriate methodological rules for regularity-refining in economics? And what is the relationship between regularity-refining and the identification of causal mechanisms or explanations? These are not idle questions, as most of the methodological reflection of economists has concentrated on theory testing.

Despite this, there is a long tradition in economics of constructing theories to explain nonexperimental empirical regularities that are perceived to be robust, often termed "stylized facts." A classic example is the postwar development of theories of aggregate consumption expenditure to explain the stylized fact that the marginal propensity to consume is greater in the long run than in the short run; a more recent example is the development of real business cycle models to explain generalizations about postwar economic fluctuations, such as the failure of employment levels and real wage rates to vary inversely together over the cycle.[18]

[17] A prominent recent discussion of this type of experimental research is provided by part II of Smith (2008).

[18] The consumption function puzzle has been a textbook standard for motivating permanent income and life cycle theories of aggregate consumption for much the postwar period. For its continuing use in this role, see, for example, Mankiw (2007, chapter 16). For surveys of real business cycle theory, see, for example, Plosser (1989), Stadler (1994), and King and Rebelo (1999).

Typically, theories constructed to explain stylized facts have implications that go beyond formal representation of those facts themselves. (There is an important tradition in the philosophy of science that *requires* this, if the explanation is to be judged a good one, as we discuss in chapter 3.) The role of experiments in *testing* such implications raises the same issues as the experimental testing of theories constructed in other ways, so that our discussion in chapters 2–3 also applies to testing theories that are constructed to explain stylized facts. But chapters 4 and 5 discuss a further question about experiments, arising specifically in relation to the formulation of theories to explain stylized facts. This question is whether the laboratory can provide stylized facts worth explaining.

On the one hand, the control that experiments offer seems to lend them well to a task of regularity-refinement. This may be hard to achieve with field data, where there can be a surprising degree of ambiguity, not always initially apparent, about what the stylized facts actually are.[19] But, conversely, just as a skeptic might suspect that experiments are too remote from the intended domain of theories to provide a useful testing ground for them, she might also suspect that experiments are too artificial to generate stylized facts that can be expected to hold outside the laboratory.

Concerns about whether experimental findings are reliable guides to what may happen outside the laboratory are sometimes expressed as doubts about their *external validity*. The question of whether conclusions reached on the basis of observations in one sphere are informative about what may happen in another sphere is not specific to experiments. It is a potential concern for any empirical research. But the artificiality of the laboratory arguably sharpens the question for experimental research.

Chapter 5 considers different senses in which the laboratory might be taken to be artificial and discusses their implications for external validity, in the contexts of both theory testing and regularity-refinement. The illustrative experiments described in section 1.3 vary considerably in how far the issue of artificiality seems prima facie to arise. For example, Bryan and Test did not conduct the experiments reported in their 1967 paper in a laboratory at all, if the latter is construed as a physical location dedicated to the experimenter's purpose. They went to considerable lengths to conceal from participants the fact that they were the subjects

[19] The cases of the consumption function puzzle and the cyclical properties of real wage rates illustrate this. See Stock (1988) on the consumption function puzzle and Abraham and Haltiwanger (1995) on the difficulties of measuring cyclical properties of real wage rates.

of an experiment, by intervening in a disguised way in a naturally occurring environment. This is an example of what, much later, Harrison and List (2004, p. 1,014) would describe as a "natural field experiment." While experiments of this naturalistic kind may mitigate concerns about artificiality, they may also give rise to other concerns. For example, Bryan and Test conducted their flat tire experiment in the street in Los Angeles, an environment in which it would not have been possible for them to hold constant everything except the treatment.[20] In contrast, Ball et al.'s (2001) investigation of the effect of status on the distribution of gains from trade in market transactions was conducted in laboratory conditions and may have a stronger claim to have held everything constant across treatments, other than the treatment manipulations themselves, because (as is normal in laboratory experiments) they assigned subjects to treatments at random. However, some, such as Bardsley (2005), have questioned whether their treatment manipulations really confer status, in the same sense as agents in the field may have it. A possible view of the comparison between Bryan and Test (1967) and Ball et al. (2001) is that it exemplifies a trade-off between naturalism and external validity, on which Bryan and Test score highly, and control, which Ball et al. (2001) have to a greater degree.[21] Early economic experimenters emphasized the virtues of control but the recent wave of field experiments in economics may reflect an increasing premium now put by some on external validity (see, for example, Levitt and List 2007). Chapter 5 considers the nature and implications of any such trade-off.

Chapter 5 also focuses on a common practice in experimental economics: namely, that of using designs that closely implement the assumptions of particular theoretical models. Among our examples there is considerable variation in how far this practice is followed. On the one hand, Bryan and Test (1967) make no attempt to implement a model; on the other hand, Goeree and Holt (2001) have subjects play games which, if they were motivated to maximize their own monetary payoffs,

[20] It is, nevertheless, remarkable how much control Bryan and Test were able to achieve. For example, a possible source of uncontrolled variation is how much of a hurry drivers were in, as that might vary over the course of the day. As noted earlier, Bryan and Test attempted to eliminate any effect of this on their results by repeating the experiment on another Saturday, with the timings of the two treatments reversed. But this clever idea still cannot counteract completely the possibility that, for reasons unrelated to the experiment, drivers became more hurried through the day on one Saturday and less hurried through the day on the other.

[21] The suggestion that there is a trade-off should not be taken as implying that a given increase in naturalism is always associated with a corresponding loss of control. We have already commented favorably on how much control Bryan and Test (1967) were able to achieve; it would certainly have been possible to design equally naturalistic, but much less well controlled, experiments.

would resemble certain abstract games of game theory, not just in their structure but also their presentation; and Morgan et al. (2006) go to considerable lengths to create a setup which, except for the use of human subjects in place of firms, matches the structure of the clearing-house model of price setting.

In a theory-testing context, some might argue that a theoretical model can only legitimately be tested in a laboratory experiment that closely implements its assumptions. This position is an assertion about the domain of the theory and, for that reason, chapter 5 applies and builds on the framework presented in chapter 2 in discussing it. However, the question of whether designs should implement the assumptions of models also arises in the context of regularity-refinement. Here, it takes on a slightly different hue: if an experiment closely implements the assumptions of some theoretical model of a particular real-world phenomenon, would it be legitimate to read its findings as establishing stylized facts about that phenomenon? This is a question of external validity and, once again, it connects with a type of artificiality of the laboratory. Because of the abstraction used in economic models, it is often the case that the more closely a design implements a formal model, the more artificial the experimental environment defined by it seems.

In chapters 6 and 7, we turn to two issues of everyday concern to experimenters: incentives, and interpretation of data in the light of stochastic variation. We suggest that experimental economists have been too prone to lapse, in the first case, into unreflective conformism, and, in the second case, into unreflective diversity. (Of course, we do not make either charge against all experimental economists. To name just two papers concerned with these topics, Camerer and Hogarth (1999) discuss incentives and Hey and Orme (1994) stochastic modeling of data from choice experiments.)

The issue of incentives is sometimes seen as dividing experimental economics from experimental psychology (Hertwig and Ortmann 2001). Reflecting this view, all those experiments among our examples that were conducted by economists presented subjects with a situation in which they could receive a sum of money determined by their own task responses in conjunction with those of other subjects and/or with chance, according to preset rules devised by the experimenter. Thus, for economists, it seems that creating rules to determine task-related monetary payoffs is a core part of experimental design; it is part of the process of creating a controlled laboratory environment. In contrast, the psychologists Tversky and Kahneman make no attempt to restrict their investigations of framing to cases where real outcomes hang on their choices. In the Asian disease problem from Tversky and Kahneman (1981), subjects

were simply asked to say which option they would "favor" in an imaginary scenario. For the unconscious subjects of the experiments reported in Bryan and Test (1967), real consequences hung on their actions (either they changed the tire or they did not; either they put money in the collection box or they did not) but the experimenters simply relied on whatever incentives the naturalistic setting provided.

However, the differences between the designs used by economists and psychologists may not always be as great as it might seem. In any experiment, subjects arrive with whatever motivations they arrive with and face the decision problems set up by the experimenter. In general terms, the fact that, even when the experimenter seeks to induce certain preferences on the part of subjects, they may still have traces of "homemade" preferences, is a reflection of the Duhem–Quine problem discussed in chapter 3; and whether economic theory should only be expected to apply in situations where substantial sums of money are at stake or the profit motive holds sway is part of the question of the domain of the theory discussed in chapter 2. But chapter 6 goes beyond these earlier discussions to look in more detail at certain issues.

One set of questions concerns whether, or maybe better, when or at what level, task-related incentives are required by good practice. A second set of questions, on which the answers to the first set may partly depend, is what effect task-related incentives actually have on subjects' behavior in the laboratory and why. We have already seen from our examples that there are differences between disciplines on whether task-related incentives are viewed as required, but, even within experimental economics, there are also differences of opinion about their level and about whether (or, perhaps again better, when) behavior is sensitive to this. More fundamentally, why do incentives affect behavior in the laboratory? A natural first thought for economists is that stronger incentives stimulate greater effort and so better performance by subjects. But, as chapter 6 discusses, this view, though fruitful, still raises a number of issues and is, in any case, only one among several perspectives on why incentives might matter for experimental economics.

A further set of questions, considered in the second half of chapter 6, arises because certain tasks that experimenters have wanted to present to subjects are not straightforward to incentivize. Does this mean that such tasks should be avoided? A related, but perhaps more common, issue in experimental economics arises because, for many tasks, it is straightforward to link subjects' rewards to what happens in the experiment in some way or other, but the more obvious ways of doing so do not satisfy orthodox economic theory. The question is then whether

*incentive compatibility* in the eyes of orthodox theory should be seen as a requirement.

To motivate this issue, consider the study reported in Selten et al. (1999) concerning the binary lottery incentive system. As noted earlier, this device is sometimes used by experimenters who wish to investigate game play, while excluding the possibility that the behavior they observe is due to risk aversion. At first sight, it may seem easy to implement, say, the payoff matrix of a particular game by paying subjects sums of money proportional to the payoffs specified by the matrix for the particular combination of choices that they have made. However, formal game theory typically works with descriptions of games in which payoffs are utilities, in the sense of Von Neumann and Morgenstern, not sums of money. These utilities are supposed to capture players' attitudes to risk; and expected utility theory specifies that they are only linear in money when the agent is risk neutral for monetary risks. So, according to orthodox theory, paying monetary amounts proportional to the payoffs of a given payoff matrix does *not* implement the relevant game if, in fact, subjects are risk loving or risk averse. The binary lottery system is an attempt to circumvent this problem, motivated by the fact that, according to expected utility theory, a subject's expected utility is linear in *chances of winning* some given sum of money, even when it is not linear in *money* itself. But the findings of Selten et al. (1999) cast doubt on whether this technique works in practice. They provide an early example of one of the themes of chapter 6; namely that incentive systems that "work" according to conventional economic theory may have unexpected effects in practice. This does not mean that such systems should never be used—for example, in testing a theory it is legitimate to use an incentive system that is valid according to that theory, even if the theory turns out to fail—but it does raise the question of how high a premium should be set on incentive compatibility, in the sense that economists have usually conceived it, and suggest that questions about the structure of incentives for an experiment cannot be settled in the abstract. They depend on the purpose at hand.

Finally, all of the example studies that we have presented have one thing in common: they draw conclusions from experimental data. The question therefore arises of how to assess the reliability of such conclusions. In a broad sense, almost the whole of this book is concerned with some aspect of this question. However, there is a further issue which, although related to the themes of chapters 2–6, is not their main focus, and yet is likely to occur to many nonexperimental applied economists. This is the issue of stochastic specification; we consider it in chapter 7.

In nonexperimental empirical economics, issues of stochastic specification and model selection loom large, but this is less common in experimental papers. This may reflect a tendency for experimenters to think that, given a well-controlled experimental design, the data will speak for themselves. But is this always correct? Or might it be argued that, just as many mid-twentieth-century economists were wrong to think that the methods of empirical economics were a necessary adaptation to the unavailability of laboratory techniques, so some late-twentieth-century experimenters have been wrong to think that their use of such techniques renders attention to stochastic specification unnecessary? Even if laboratory techniques diminish or exclude some of the sources of stochastic disturbance in field data, it does not follow that they do so for all sources.

To motivate this issue in relation to theory testing, note that most economic theories are essentially deterministic. However, since the early development of econometrics,[22] applied economists have recognized that such models cannot be literal descriptions of the processes that generate the data typically studied because, relative to this standpoint, there is too much variability in the data. This is as true of experimental data as it is of field data. Indeed, arguably, it can be seen even more clearly in experiments than in the field, because the laboratory allows us to face the same subjects with essentially the same task on more than one occasion, with almost everything held constant except for the passage of time.

Part of the process of confronting a theory with data therefore involves stochastic specification of the model. This would be unproblematic if there were a clearly correct, all-purpose method of stochastic specification that could be applied to deterministic theories. But this is not so; and, worse, different methods may lead to different conclusions about a particular theory from a given set of experimental findings—another instance of the Duhem–Quine problem from chapter 3. Further, it may not be the case that employing routine econometric techniques when analyzing experimental data, without seriously considering the sources of variability in that data, will be helpful. We illustrate and discuss these problems in chapter 7, comparing and assessing alternative types of stochastic specification for economic theories of choice and strategic interaction. Our earlier discussion of the findings of tests of these theories is, to some extent, conditional on this later treatment. Chapter 7 also considers whether different branches of experimental economics

[22] See Morgan (1990) for a history of econometrics.

have something to learn from each other on the question of stochastic specification and extends the argument introduced in chapter 6 that experimenters should not be shy of new forms of experimental data.

Chapter 8 concludes by drawing the themes of the book together and, as a consequence, suggesting that a number of popular nostrums concerning experimental economics are misplaced.

## 1.5 Methods, Methodology, and Philosophy of Science

At the start of this chapter, we described this book as a *methodological* assessment of experimental economics. What does that mean? According to its dictionary definition, "methodology" is the systematic study of method, and, by extension, method considered in a systematic way. So, apart from sounding grander, is the methodology of experimental economics anything more than the body of methods that experimental economists use?

Professional methodologists sometimes distinguish between "methods" and "methodology," conceiving the former as part of the routine practice of a science and the latter as higher-order reflection. In a recent survey of economic methodology, Wade Hands (2001) defines "methods" as "the practical techniques employed by successful economists in the execution of their day-to-day professional activities." Methodology in the sense of "methods," he says,

> is essential to professional success, usually acquired tacitly, or by rote, in the context of actually working on specific economic research projects: initially under the guidance of one's research supervisor or thesis director, and then later through interactions with one's colleagues, department chair, and various journal editors. It is the source of answers to day-to-day questions like: Is an $R^2$ this low OK for this kind of model? Is it reasonable to assume the Jacobian matrix has this strange sign pattern? or, It's OK to drop all of the data from the first two quarters of 1929, right? As important as such questions might be, [method in this sense] is *not* what most economists mean when they use the term Economic Methodology. [This] is not generally what one will see published in journals like *Economics and Philosophy* or *The Journal of Economic Methodology*, which specialize in methodological research; and, whereas one might overhear such topics discussed by Nobel laureates, it is not what they write about when they write about "Methodology."
>
> Hands (2001, p. 3)

The suggestion seems to be that "method" is a set of relatively uncontroversial rules of good practice, internal to a scientific discipline, of which established scientists have a tacit understanding and into which

novices are inducted. "Methodology" is a more elevated or abstract activity, pursued by professional methodologists (and, apparently, by Nobel laureates in their more reflective writings).

Clearly, there is a difference of some kind between the questions that Hands classifies as "method" and those that professional methodologists discuss. Still, we would have difficulty in classifying our subject matter either as "method" or "methodology," in Hands's terms. In relation to experimental economics, at least, we think this distinction is unhelpful.

Notice how Hands's examples of the processes by which "method" is learned involve scientists passing judgment on the scientific quality of other scientists' work. The research student's thesis is judged by the supervisor or examiner; the author's research paper is judged by the journal editor; the research record of the applicant for the academic post is judged by the department chair; the standing of the researcher in his field is judged by his colleagues. In all of these processes, judgments about what does and does not constitute good science are invoked. It is in these "day-to-day" processes that the battles which determine how the accepted methods of a science evolve are fought. If a new method is to establish its legitimacy, the scientists who favor it have to convince other scientists of its merits: research students have to convince skeptical supervisors or examiners, authors have to convince skeptical referees and editors, job applicants have to convince skeptical department chairs, colleagues have to convince skeptical colleagues. All of this has certainly been true of the introduction of experimental methods into economics: we, the authors, have taken part in these kinds of battles.

It is only after many day-to-day judgments have been made in favor of a new method that it is likely to come to the attention of professional methodologists or its pioneers to attain the status of Nobel laureates. These judgments involve more than the routine application of accepted rules of good practice, even though neither the judges nor the judged typically appeal explicitly to the higher-order principles of philosophy of science. Much of our book will be concerned with judgments of this kind. That is, we will be concerned with controversies *among practicing economists* about what the methods of their discipline should be. Or, since we are practicing economists addressing practicing economists, we can say: what the methods of *our* discipline should be.

A methodological assessment of a new development in a discipline must, we maintain, engage with issues of "method." But it must also be informed by "methodology" in its higher-order sense; methodology as practiced by professional methodologists. In the remainder of this section, we explain how our approach to our subject matter relates to

methodology, so understood. We begin with a thumbnail sketch of what methodologists do.

Traditionally, methodological enquiry begins in philosophy of science or, more fundamentally, in *epistemology*—that branch of philosophy which studies the nature of knowledge and justified belief. Epistemology addresses questions such as, What are the defining conditions of knowledge? What are the substantive sources of knowledge? Does it have foundations, and if so, what are they? Are there limits to what can be known? The mainstream of work in philosophy of science has investigated the special characteristics of scientific knowledge, often implicitly presupposing that the received theories of natural science have especially strong claims to the status of knowledge.

For much of the twentieth century, philosophy of science was dominated by the search for a satisfactory *empiricist* account of knowledge: the aim was to identify general principles by which knowledge can be generated *from observation and experiment*. There was a presumption that these principles would be implicit in the best practices of the most "advanced" natural sciences, particularly physics. By identifying those principles, it would be possible to distinguish between science and non-science; some disciplines that claimed to be sciences might then be revealed to be merely pseudosciences.

Perhaps because of the deeply rooted aspiration of economists to the status of practitioners of "hard" science, this program of philosophical investigation has had a particularly significant impact on economics. It has affected both the methods used in the discipline and economists' interpretations of them. That influence can still be seen in the language of practical economists. For example, the idea that there is a categorical distinction between "descriptive" (or "positive") propositions on one side and "normative" propositions (or "value judgments") on the other, and that economics as a science is concerned only with the former, entered economics from a particular strand of empiricist philosophy of science. So too did the idea that a theory has content only if it leads to hypotheses that can be tested against observations—an idea that has led generations of economists to append a list of testable hypotheses to every theoretical proposal. Economics has been particularly influenced by the empiricist philosophy of Karl Popper (1934), known as *falsificationism*, which we will discuss in relation to experimental economics in chapter 3. Popper's central idea is that scientific knowledge consists of a body of hypotheses that in principle are capable of being falsified by observation but in fact have withstood the best efforts of the scientific community to find disconfirming evidence. On this account, scientific virtue is identified with *bold conjectures*: that is, putting forward hypotheses which, a priori,

appear improbable and which lay themselves open to many possibilities for falsification—and *severe tests*: that is, subjecting hypotheses to tests which, a priori, seem particularly likely to falsify them. Bold conjectures that survive severe tests are ones in which confidence can be placed.

Over the last two decades, issues of epistemology have become less central to philosophy of science. Perhaps this reflects a recognition of the problems posed for empiricism and positivism by the Duhem–Quine problem, but it might simply be a change in intellectual fashion. Whatever the reason, there has been a shift of emphasis in philosophy of science from issues of epistemology to issues of *ontology*—that branch of philosophy that studies the nature of existence. In particular, philosophers of science have proposed and discussed various theories of *realism*. In traditional empiricist accounts, knowledge is grounded on observation. Ultimately, knowledge is *about* observations; the problem for science is to find regularities in past observations that can be used to make reliable predictions about future observations. In realist accounts, in contrast, our observations can inform us about *forces* or *capacities* or *mechanisms*; those mechanisms, even if not straightforwardly observable, cause the regularities we observe.

The realist project can be pursued in different ways. Some realist philosophers—for example, Richard Boyd (1983)—start from (what they take to be) the fact that the natural sciences have been extraordinarily successful in predicting observable features of the world, and in showing us how to manipulate the world to produce chosen results. Those predictions and manipulations are based on theories that (it is claimed) postulate unobservable mechanisms. The most credible explanation of the success of science, it is then argued, is that the mechanisms it postulates, or mechanisms very like them, really exist. This brand of realism asks what properties the world would have to have in order for science as we know it to be successful, and then infers the existence of those properties from the success of science. A different approach, which does not presuppose the success of science, is to ask what properties have to be attributed to the world in order for scientists' *claims* to knowledge to make sense on their own terms. If we can make sense of those claims only on the supposition that scientists are postulating the existence of real mechanisms, then we can conclude that science *is committed to* realism. This argument is developed by Nancy Cartwright (1989) in relation to the practices of physics and economics.[23] A third approach, pursued

[23] Cartwright (1989, p. 158) argues that, in view of the success of physics, the capacities postulated by physics "are scarcely to be rejected." But she is explicitly agnostic about whether economics is a successful science, claiming only that economics is committed to realism.

by Tony Lawson (1997), turns Boyd's "argument from success" on its head. If a science is consistently *un*successful, perhaps the most credible explanation is that the mechanisms that it postulates do *not* exist. Thus, according to Lawson, one way to set about the reconstruction of an unsuccessful science is to start in ontology, by proposing new conceptions of the sorts of mechanisms that really operate beneath the surface of the phenomena that are to be explained. Lawson maintains that "orthodox" economics *has* failed, and recommends economists to start all over again with his preferred form of "critical realism."

As a first step in explaining how our book fits into the grand scheme of methodology, we declare that, in our roles as its authors, we do not set out to defend any particular philosophical position in either epistemology or ontology. We are not looking for foundational principles from which we can construct rules of good practice in experimental economics. Nor are we primarily concerned with uncovering ontological assumptions implied by current practice.

To illustrate our stance toward fundamental methodological issues, we return to the debate between empiricist and realist conceptions of scientific knowledge. Within this debate, interpretations of the concepts of "explanation" and "causation" are hotly contested. For realists, there is a categorical distinction between *observing* a regularity in the world and *explaining* it; an explanation appeals to a supposed causal mechanism, while an observation is just an observation. For empiricists, in contrast, science is simply about finding regularities in observations: the only sense in which some regularity can be "explained" is by showing it to be an instance of a more general regularity. Similarly, talk about "causation" is to be understood as just another way of referring to observed regularities.[24] But while methodologists argue about what scientists *really* mean, or *ought to* mean, by "explanation" and "causation," practicing scientists are generally able to use these concepts in a mutually intelligible way without getting ensnared in ontological debate. Our experience of economics suggests to us that this is not because scientists are all realists, or because they are all empiricists. It is because much of what they actually need to say about explanation and causation would be just the same whichever of these positions they maintained. In writing this book, our standpoint is that of practicing experimental economists. Our default position is that we too can write intelligibly about explanation and causation without committing ourselves to ontological assumptions.

[24] The classic empiricist account of causation is that of Hume (1739–40, see 1978 edition, pp. 155–72). Hume argues that causation is a property not of the external world, but of our mental perceptions. The perception of causation is a psychological response to the observation of certain kinds of regularity.

In the rest of this book, the distinction between empiricism and realism will appear if, but only if, it matters for the arguments we want to make.

More generally, our understanding of the relationship between methodology and methods can be expressed in the following way. Without claiming deep insight into the abstract reaches of philosophy of science, we declare our sympathies with the anti-foundational epistemology of Willard Van Orman Quine (1951). By this, we do not mean that we will be presenting arguments in support of Quine's philosophy, but only that our approach to our subject matter is broadly Quinean in spirit. For Quine, the idea of finding the *foundations* of scientific knowledge is misguided. No form of belief, not even in the reality of "raw" observations or in the supposedly "analytic" (that is, necessarily true) theorems of mathematics and logic, is totally secure, independent of support from other beliefs. In place of the metaphor of foundations, Quine offers that of the *web of belief*:

> The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges with experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field.…the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience. No particular experiences are linked with any particular statements of the interior of the field, except indirectly through considerations of equilibrium affecting the field as a whole.
>
> Quine (1951, section 6)

Another vivid metaphor, with much the same meaning as Quine's, is the ship imagined by Otto Neurath (1937):

> We possess no fixed point which may be made the fulcrum for moving the earth; and in like manner we have no absolutely firm ground upon which to establish the sciences. Our actual situation is as if we were on board ship on an open sea and were required to change various parts of the ship during the voyage.
>
> Neurath (1937, p. 276)

The idea behind these metaphors is that the body of scientific knowledge has no foundations: every part of it relies on other parts for corroboration. If we think of the study of methodology in these terms, we have to recognize that what we can learn from the philosophy of science is neither more nor less "fundamental" than what we can learn from the practice of science: methodology is neither more nor less fundamental than

methods. If abstract philosophical reasoning can generate particular conclusions about the nature of knowledge, and if those conclusions imply that particular scientific methods are more likely than others to generate reliable knowledge, then that provides some support for whatever knowledge claims are produced by the favored methods. But, conversely, if the application of particular scientific methods is found to produce successful predictions and to assist us in manipulating the world, then that provides some support for whatever lines of philosophical reasoning lead to the conclusion that those methods are reliable.

This general conception of the mutual dependence of methodology and method is compatible with much of the current practice of methodology. In particular, it is compatible with the reluctance of many modern methodologists to prescribe regulative principles of scientific method in the way that, for example, Popper saw it as his job to do, and with modern methodologists' interest in investigating how, in particular sciences, claims to knowledge are in fact established and contested.[25] This less regulative approach is sometimes presented as showing decent humility on the part of the methodologist. Is it not presumptuous for philosophers of science to claim to judge the validity of the methods used by practicing scientists? If we accept Quine's metaphor of the web of belief, it is natural to ask, In which part of the web do we have greater confidence—in the substantive claims of received scientific theories or in the philosophical claims of epistemologists? If the former, should we not respond to conflicts between the two by reevaluating our epistemology rather than by questioning the methods used by scientists?

A significant example of this argument can be found in the work of Thomas Kuhn. Kuhn has had a huge impact on methodology through his historical study of "scientific revolutions." According to Kuhn's account, a "paradigm"—an assemblage of cohering theories, questions, and practices—can continue to hold the allegiance of a scientific community in the face of what later comes to be recognized as a mass of disconfirming evidence. The overthrow of a paradigm is a social process as much as it is the systematic application of rules of scientific method (Kuhn 1962). In the eyes of some critics, Kuhn's uncritical acceptance of this alleged characteristic of science is irrational and relativist. Kuhn makes the following reply to these criticisms:

> To say that, in matters of theory-choice, the force of logic and observation cannot in principle be compelling is neither to discard logic and observation nor to suggest that there are not good reasons for favoring

[25] Cartwright's (1989) study of knowledge claims in econometrics is an example.

> one theory over another. To say that trained scientists are, in such matters, the highest court of appeal is neither to defend mob rule nor to suggest that scientists could have decided to accept any theory at all.
>
> Kuhn (1970, p. 234)

In other words, Kuhn is not denying that there are standards of good scientific method; but he *is* denying that philosophy of science is qualified to define or police those standards.

The development of experimental economics over the last few decades has some similarities with a revolution—or, at least, an attempted revolution—in Kuhn's sense. Previously accepted ideas about the proper methods of the discipline and the criteria that should be used to choose between rival theories are being fought over. The implication of Kuhn's argument is that philosophy of science should not be expected to provide metacriteria by which to adjudicate whether the cause of true science is represented by the revolutionaries or the ancien régime. Science has to regulate itself.

While this argument has force for someone who is thinking about science from outside, it cannot provide the structure for our work. In Kuhn's sense, we *are* trained scientists. With respect to issues of method and theory-choice in economics, we are *members of* Kuhn's "highest court of appeal." Our book is not a commentary on the judgments of this court; it is part of the judgment process itself. In saying this, we do not lay claim to any special authority. Rather, the metaphor of courts of appeal is not completely apt. Courts of justice are organized hierarchically; one appeals from lower to higher ones. In contrast, science has more of the nature of a spontaneous order. A knowledge claim in economics becomes accepted as the cumulative effect of many small acts of judgment. Notwithstanding the existence of major players, there is no supreme court, no committee of the great and the good of the discipline, which collectively certifies or overrules the outcomes of this process.

This naturally raises the questions, What status can we claim for our judgments? On what are they grounded? How can the reader test their validity?

Before answering these questions directly, we take a step back. Recall Hands's examples of the "day-to-day" contexts in which methods are learned, and which we redescribed as potential locations of contests about which methods should be preferred to which. These are situations in which judgments are made about knowledge claims. How are these judgments expressed? Think of a panel of examiners judging a dissertation, an editor and her referees judging a paper, an appointment committee judging a job applicant, a panel of distinguished scholars

awarding a major prize. The reader may have noticed that, in each case, we have replaced the single judge from Hands's example with a *panel* of judges. This revision reflects the reality of scientific institutions: collective judgments are the norm. If the members of a panel disagree—if, for example, one referee thinks a paper should be published while another thinks it should be rejected—there is an expectation that each will provide *reasons* for his judgment, in a form which allows discussion about its merits. And even if a panel is in agreement, there is often an expectation (as in the case of rejection decisions by journals) that the agreed judgment is supported by reasons that can be communicated to the person whose work is being judged. According to the conventions of refereeing, it is not enough just to say: "I am a trained economist, with a feel for the tacit rules of economics. I have a sense, which I cannot articulate, that this paper is no good."

The point of these examples is that, as a matter of simple fact, scientific practice involves not only the making of judgments between alternative methods and between competing theories; it also involves the giving of reasons. And more than that: reasons can be judged to be good or bad, sound or unsound, strong or weak. For example, if two members of an editorial board disagree about the suitability of a particular paper for publication, each may challenge the validity or force of the reasons that the other has given for her judgment. And again, it is not enough for one just to say to the other: "I am a trained economist, with a feel for the tacit rules of argument within the discipline. I have a sense, which I cannot articulate, that your reasons are no good." Further reasons are expected.

At first sight, it might seem that this chain of reasons and metareasons must *either* lead back to the ultimate foundations of knowledge (if such things exist), *or* lead back to claims for which no reasons can be given, *or* lead to an infinite regress. But, when individuals are trying to resolve a disagreement, there is no need for them to go back further than propositions on which they agree. Since these propositions are not in dispute, reasons are not called for. In disputes between practicing scientists, the search for reasons will often stop well before the issues that concern professional methodologists are reached. For example, the two members of our editorial board might resolve their disagreement about the merits of the paper without ever becoming aware that one of them believes that scientific knowledge can only be about observation while the other believes that science can discover real but unobservable mechanisms. Sometimes, however, the parties to what appears to be a day-to-day disagreement may find that they are disagreeing about issues in philosophy of science. And then, contrary to the implication of Kuhn's metaphor of

the highest court of appeal, issues of method may be resolved by appeal to principles of methodology. Our approach will allow both possibilities. We will neither start from professional methodology or philosophy of science, nor shy away from them if that is where the arguments lead us.

We can now answer the rhetorical question we asked a few paragraphs back, What status can we, the authors of this book, claim for our judgments? All we can claim is that we are contributing to ongoing debates within economics, following what we believe to be the implicit rules of engagement. Our aim is to support our judgments by appealing to premises that our readers will accept and by giving reasons that our readers will find convincing.