

Markets, Games, and Strategic Behavior: Recipes for Interactive Learning

Charles A. Holt
University of Virginia
Comments welcome: holt@virginia.edu

February 20, 2006

Markets, Games, and Strategic Behavior – Charles A. Holt

Charles A. Holt
University of Virginia

© Copyright, All Rights Reserved
All rights reserved.

Chapter 1. Introduction

I. Origins

Like other scientists, economists observe naturally occurring data patterns and then try to construct explanations. Then the resulting theories are evaluated in terms of factors like plausibility, generality, and predictive success. As with other sciences, it is often difficult to sort out cause and effect when many factors are changing at the same time. Thus, there may be several reasonable theories that are roughly consistent with the same observations. Without a laboratory to control for extraneous factors, economists often “test” their theories by gauging reactions of colleagues (Keynes, 1936). In such an environment, theories may gain support on the basis of mathematical elegance, persuasion, and focal events in economic history like the Great Depression. Theories may fall from fashion, but the absence of sharp empirical tests leaves an unsettling clutter of plausible alternatives. For example, economists are fond of using the word equilibrium preceded by a juicy adjective (e.g. proper, perfect, divine, or universally divine). This clutter is often not apparent in refined textbook presentations.

The development of sophisticated econometric methods has added an important discipline to the process of devising and evaluating theoretical models. Nevertheless, any statistical analysis of naturally occurring economic data is typically based on a host of auxiliary assumptions. Economics has only recently moved in the direction of becoming an experimental science in the sense that key theories and policy recommendations are suspect if they cannot provide intended results in controlled laboratory and field experiments. This book provides an introduction to the study of economic behavior, organized around games and markets that can be implemented in class.

The first classroom market games were conducted in a Harvard class by Edward Chamberlin (1948). He had proposed a new theory of “monopolistic” competition, and he used experiments to highlight failures of the standard model of perfect competition. Students were given buyer and seller roles and instructions about how trades could be arranged. For example, a seller would be given a card with a “cost” in dollars. If the seller were to find a buyer who agreed to pay a price above this cost, the seller would earn the difference. Similarly, a buyer would be given a card with a “resale value,” and the buyer could earn the difference if a purchase could be arranged at a price below this resale value. Different sellers may be given cards with different cost numbers, and likewise, buyers may receive different values. These values and costs are the key elements that any theory of market price determination would use to derive predictions, as explained in Chapter 2. Without going into detail, it should be clear that it is possible to set up a laboratory market, and to provide financial incentives by using

value and cost numbers in dollar amounts and by paying subjects their earnings in cash.

Chamberlin's classroom markets produced some inefficiencies, which he attributed to the tendency for buyers and sellers in a market to break off and negotiate in small groups. Vernon Smith, who attended Chamberlin's class, later began running classroom markets with an enforced central clearinghouse for all offers to buy and sell. This trading institution is called a "double auction" since sellers' ask prices tend to decline at the same time as buyers' bid prices rise. A trade occurs when the bid-ask spread closes and someone accepts another's offer to buy or sell. Smith observed efficient competitive outcomes, even with as few as 6-10 traders. This result was significant, since the classical "large numbers" assumptions were not realistic approximations for most market settings. Smith's early work on the double auction market figured prominently in his 2002 Nobel Prize in Economics.

A parallel development is based on game-theoretic models of strategic interactions. In a "matching-pennies" game, for example, each player chooses heads or tails with the prior knowledge that one will win a sum of money when the coins match, and the other will win when the coins do not match. Each person's optimal decision in such a situation depends on what the other player is expected to do. The systematic study of such situations began with John von Neumann and Oscar Morgenstern's (1944) *Theory of Games and Economic Behavior*. They asserted that standard economic theory of competitive markets did not apply to the bilateral and small-group interactions that make up a significant part of economic activity. Their "solution" was incomplete, except for the case of "zero-sum" games in which one person's loss is another's gain. While the zero-sum assumption may apply to some extremely competitive situations, like sports contests or matching pennies games, it does not apply to situations where all players might prefer some outcomes to others.

Economists and mathematicians at the RAND Corporation in Santa Monica, California began trying to apply game-theoretic reasoning to military tactics at the dawn of the Cold War. In many nuclear scenarios, it is easy to imagine that the "winner" may be much worse off than would be the case in the absence of war, which results in a non-zero-sum game. At about this time, a young graduate student at Princeton entered von Neumann's office with a notion of equilibrium that applies to a wide class of games, including the special case of those that satisfy the zero-sum property. John Nash's notion of equilibrium and the half-page proof that it generally exists were recognized by the Nobel Prize committee about 50 years later. With the Nash equilibrium as its keystone, game theory has recently achieved the central role that von Neumann and Morgenstern envisioned. Indeed, with the exception of supply and demand, the "Nash equilibrium" is probably used as often today as any other construct in economics.

Intuitively speaking, a Nash equilibrium is a set of strategies, one for each player, with the property that nobody would wish to deviate from their planned action given the strategies being used by the other players. For example, consider a situation in which each person would prefer to go to work if the other person does, but would prefer to stay home otherwise, since working alone is less productive in this case. In this “coordination game,” there could be two equilibria, one in which both work and another in which the both stay home. This game could be implemented in class by letting each player choose one of two playing cards, with each card corresponding to one of the two decisions.

These classroom markets and games can be quite useful, because participants learn what an economic situation is like “from the inside” *before* seeing standard presentations of the relevant economic theory. A classroom game (followed by structured question-and-answer discussion) can let participants discover the relevant economic principle for themselves, which enhances the credibility of seemingly abstract economic models. Each of the chapters that follow will typically be built around a game or market that can be run in class, using a web-based software suite and/or simple props like dice and playing cards.

II. Overview

Game Theory

A careful analysis of a strategic situation typically involves more than just identifying an equilibrium. For example, an equilibrium may be a bad position from which deviations are not so costly for any individual. Alternatively, there may be multiple equilibria, so that standard equilibrium theory makes no prediction. Observed behavioral tendencies in experiments can provide important guideposts for the development of new theoretical approaches. Indeed, the mathematicians and economists at the RAND Corporation began running game experiments at about the same time as Chamberlin’s classroom market experiments. Basic game-theoretic concepts are introduced in Chapters 3 and 5, and the behavioral game theory chapters in part VI provide a careful consideration of conditions under which behavior does or does not conform to equilibrium predictions.

Individual Decisions

A game or market may involve a relatively complex set of interactions between multiple players or traders. Sometimes it is useful to study key aspects of behavior of individuals in isolation. For example, stock markets involve major risks of gains and losses, and it is instructive to consider how individuals react to simple risks that are not generated by the behavior of others in complicated market settings. It is straightforward to set up a simple decision experiment by

giving a person a choice between gambles or “lotteries,” e.g. between a sure \$1 and a coin flip that yields \$3 in the event of heads. Which would you choose in this case? What if the choice were between a sure \$100,000 and a coin flip that provides a 50-50 chance of \$0 and \$300,000? These types of decisions are considered in Chapter 4, which introduces the notions of expected values and risk aversion. Basic notions of decision-making under risk are useful in the analysis of interactive situations in subsequent chapters. In addition, the chapters in part VII of the book are focused on a series of specific decision-making situations, e.g. prediction, search, and information processing.

Markets

The chapters in part II cover several ways that economists have modeled market interactions, with firms choosing prices, production quantities, quality grades, or entry decisions. Some markets have distinct groups of buyers and sellers, and others more closely resemble stock markets in which purchase and resale is common. Market experiments can be used to assess the antitrust implications of mergers, contracts, and other market conditions. One goal of such experiments is to identify factors that increase the extent to which markets achieve all possible gains from trade, i.e. the *efficiency* of the market. Laboratory markets with enough price flexibility and good information about going prices tend to be highly efficient. In some contexts, however, laboratory markets can fail to generate efficient outcomes, due to imperfections in information, market power, or collusion among sellers.

Bargaining

Economic decisions in small-group settings often raise issues of fairness since earnings may be inequitable. In a simple “ultimatum bargaining game,” one person proposes a way to divide a fixed amount of money, say \$100, and the other may either agree to the division, which is implemented, or may reject the division, which results in zero earnings for both. Attitudes about inequity are more difficult to model than the simpler selfish money-seeking motives that may dominate in impersonal market situations. In this case, research experiments can provide insights in areas where theory is silent or less developed. Many topics in Law and Economics, for example, involve bargaining (e.g. bankruptcy, pre-trial settlements). Bargaining experiments have also been used by anthropologists (Ensminger, 2004; Henrich *et al.*, 2001) to study attitudes about fairness in primitive societies in Africa and South America. The chapters in part III pertain to games where fairness, equity, and other interpersonal factors seem to matter.

Public Choice

Inefficiencies can occur when some costs and values are not reflected in prices. For example, it may be difficult to set up a market that allows public goods like national defense to be provided by decentralized contributions. Another source of potential inefficiency occurs when one person's activity has a negative impact on others' well being, as is the case with pollution or the over-use of a freely available, shared resource. Non-price allocation mechanisms often involve the dedication of real resources to lobbying. No individual contestant would want the value of their effort to exceed the value of the object being sought, but the aggregate lobbying costs for a number of individuals may be large relative to the value of the prize. The public choice chapters in part IV pertain to such situations; topics include voluntary contributions, the use of a "common-pool" resource, costly lobbying, and voting-based resource allocations.

Auctions

The advent of web-based communications has greatly expanded the possibilities for setting up auctions that connect large numbers of geographically dispersed buyers and sellers. Auctions are used extensively, for example, to allocate local communications bandwidth licenses to competing firms, and experiments have been instrumental in the design and testing of such auctions, in the U.S. and in many European countries. Experiments provide government officials with confidence that new procedures will function smoothly and efficiently. In the state of Virginia, for example, officials decided to adopt an innovative "clock auction" with computer driven bid increases, after observing how the process worked in laboratory experiments conducted at George Mason University. Similarly, several experimental economists at Georgia State used experiments to devise a set of rules that were used in a Georgia auction that determined which farmers would be paid not to irrigate during a severe draught in 2001. State officials observed some of these experiments before drafting the actual auction procedures. A large-scale field trial with over a hundred bidders at 5 southwest Georgia locations preceded the actual auction, which involved about 200 farmers and \$5 million dollars in irrigation reduction payments. The auction had much of the look and feel of a laboratory experiment, with the reading of instructions, a round-by-round collection of bids, and web-based bid collection and processing. The chapters in part V pertain to such auctions, including an account of current FCC initiatives for auctioning off combinations or "packages" of broadcast licenses. These initiatives are being evaluated by experiments run in laboratories at a number of universities in the U.S. and Europe, including Caltech where the author is currently a visiting scholar.

Information

Markets may also fail to generate efficient results when prices do not convey private information. With limited information, individuals often rely on “signals” like educational credentials or ethnic background. Informational asymmetries can produce interesting patterns of conformity or “herding” that may have large effects on stock prices, hiring patterns, etc. Experiments are particularly useful in these cases, since the effect of informational disparities is to produce many Nash equilibria. In addition, laboratory and field markets can be used to aggregate information held by different individuals; these are popularly called “prediction markets.” The information chapters in the final part include experiments on Bayesian learning, herding, signaling, information aggregation, and discrimination.

III. Methodology

The behavioral insights and theoretical predictions presented in the chapters that follow will be illustrated with results from experiments run in the laboratory and the field. In order to evaluate the resulting data with a careful, skeptical eye, it is essential to understand underlying methodological considerations.

Treatment Structure

A treatment is a completely specified set of procedures, which includes instructions, incentives, rules of play, etc. Just as scientific instruments need to be calibrated, it is useful to calibrate economics experiments. *Calibration* typically involves establishing a “baseline treatment” for comparisons. For example, suppose that individuals are given sums of money that can either be invested in an “individual account” or a “group account,” where investments in the group account have a lower return to the individual, but a higher return to all group members. If the typical pattern of behavior is to invest half in each account, then this might be attributed either to the particular investment return functions used or to “going fifty-fifty” in an unfamiliar situation. In this case, a pair of treatments with differing returns to the individual account may be used. Suppose that the investment rate for the individual account is fifty percent in one treatment, which could be due to confusion or uncertainty, as indicated above. The importance of the relevant economic incentives could be established if the investment rate in the individual account falls sharply when the return for investing in the individual account is reduced.

Next, consider a market example. High prices may be attributed to small numbers of sellers or to the way in which sellers are constrained from offering private discounts to particular targeted buyers. These issues could be investigated by changing the number of sellers, holding discount opportunities constant, or by

changing the nature of allowed discounting while holding the number of sellers constant. Many economics experiments involve a *2x2 design* with treatments in each of the four cells, e.g. low numbers with discounting, low numbers with no discounting, high numbers with no discounting, and high numbers with discounting.

One common design flaw is to use a treatment structure that changes more than one factor at the same time, so that it is difficult to determine the cause of any observed change in behavior. For example, a well-known study of lottery choice compared the tendency of subjects to choose a sure amount of money with a lottery that may yield higher or lower payoffs. In one treatment, the payoffs were in the zero to \$10 range with real money, and in another treatment the payoffs were in thousands of (hypothetical) dollars. Then the author concluded that there were no incentive effects, since there was no observed difference in the tendency to choose the safe payoff. The trouble with this conclusion is that the high-payoff treatment was conducted under hypothetical conditions, so two factors were being changed: the scale of the payoffs, and whether or not they were real or hypothetical. This conclusion turns out to be questionable. For example, Holt and Laury (2002) report that scaling up hypothetical choices has little effect on the tendency to select the safer option in their experiments. Thus, choice patterns with low real payoffs do look like choice patterns with both low and high hypothetical payoffs. However, scaling up the payoffs in the real (non-hypothetical) treatments to hundreds of dollars causes a sharp increase in the tendency to choose the safer option. It is interesting to note that Holt and Laury were (correctly) criticized for presenting each subject with a low-payoff lottery choice *before* they made high-payoff choices, which can confound payoff-scale and order effects. To address this issue, Laury and Holt (2005) report an experiment in which each person only made decisions in a single payoff-scale treatment, all done in the same order, which brings us to the next topic.

“Between-Subjects” Versus “Within-Subjects” Designs

Many of the following chapters are based on a single experiment. In order to preserve time for discussion, classroom experiments typically involve a pair of treatments. One issue is whether half of the people are assigned to each treatment, which is called a *between-subjects* design. The alternative is to let each person make decisions in both treatments, which puts more people into each treatment but affords less time to complete multiple rounds of decision making in each treatment. This is called a *within-subjects* or sequential design, since behavior for a group of subjects in one treatment is compared with behavior for the same group in the other treatment. The “between subjects” and “within subjects” terminology is commonly used in psychology, where the unit of observation is typically the individual subject. Subjects in economics

experiments generally interact, so the unit of analysis is normally the group or market, where a “within” design usually involves running one treatment before another, with the same group of subjects.

Each type of design has its advantages. If behavior is slow to converge or if many observations are required to measure what is being investigated, then the parallel (between-subjects) design may be preferred since it will generate the most repeated observations per treatment in the limited time available. Moreover, subjects are only exposed to a single treatment, which avoids *sequence effects*. For example, a market experiment that lets sellers discuss prices may result in some successful collusive arrangements, with high prices that may carry over even if communication is not allowed in a second treatment. When students are asked to design a classroom experiment to be run on the others, they often come up with sequences of treatments for the same group. The most common *ex post* self-assessment is that “I wish I had used a single treatment on each half of the class so that we would not have to wonder about whether the change in behavior between treatments was due to prior experience.” (For example, the reader who looks at Figure 6.4 later in the book will see a sequence of dots that seem to evolve continuously across a change in treatment, as they decline from one theoretical prediction to another.) Even in research experiments, it is annoying to have to report and analyze decisions differently depending on where they were observed in a sequence.

Sometimes sequence effects are themselves the focus of the experiment, e.g. the issue of whether the imposition of a minimum wage causes worker aspiration wage levels to increase (Fehr, 2006). If sequence effects are not the focus, they may be avoided with the use of a between-subjects design by exposing each group of subjects to a single treatment, and comparisons can be made across groups by recruiting from the same subject pool for the different treatments.

The advantage of a sequential (within-subjects) design is that individual differences are controlled for by letting each person serve as their own control. For example, suppose that you have a group of adults, and you want to determine how much running speed is increased if people are wearing running shorts instead of blue jeans. In any group of adults, running speeds may vary by factors of 2 or 3, depending on weight, age, health, etc. In this case it would be desirable to time running speeds for each person under both conditions, with alternating treatment orders, unless the distance is so great that fatigue would cause major sequence effects. In general, a within-subjects design (two or more treatments for each group) is more appealing if there is high behavioral variability across individuals or groups, e.g. the runners, relative to the variability caused by sequencing. A between-subjects design (one treatment per group) is better when there is less variability across individuals or groups, and when there are sequence effects that cause behavior in one treatment to be influenced by what happened in an earlier

treatment. Sometimes the best choice between sequential and parallel designs is not clear, and a *sequence of experiments*, one with each method, provides a better perspective on the behavior being studied, e.g. the Holt and Laury (2002, 2005) combination of within and between designs.

Incentives

Economics experiments typically involve monetary decisions like prices, costly efforts, etc. Most economists are suspicious of results of experiments done with hypothetical incentives, and therefore real cash payments are almost always used in laboratory research. As we shall see in later chapters, sometimes incentives matter a lot and sometimes they do not matter not at all. For example, people have been shown to be more generous in offers to others when such offers are hypothetical than when generosity has a real cost, and people tend to become considerably more risk averse when the stakes are very high (hundreds of dollars for a single decision) than is the case where the stakes are hypothetical or involve only several dollars. Conversely, it is hard to imagine that scaled money payments would help serious students raise their GRE scores, and there is even some psychological evidence that money payments interfere with a child's test performance. In economics experiments, the general consensus is that money payments or other non-hypothetical incentives are necessary. This is because the underlying theoretical models have agents who are assumed to be motivated by incentives. Nevertheless, there is some evidence that scaling up payoffs does not have much of an effect in many laboratory situations (Smith and Walker, 1993). There are many documented situations in the economics and psychology literature where money incentives do not seem to have much effect, but in the absence of a widely accepted framework for identifying such situations with precision, it is usually advisable to use money incentives in laboratory economics experiments.

For purposes of teaching, it is not possible or even desirable to use monetary payments. The results of class experiments can provide a useful learning experience as long as the effects of payments in research experiments are provided when important incentive effects have been documented. Therefore, the presentations in the chapters that follow will be based on a mixture of class and research experiments. When the term "experiment" or "research experiment" is used, this will mean that all earnings were at reasonable levels and were paid in cash. The term "classroom experiment" indicates that payoffs were basically hypothetical. However, for non-market classroom experiments discussed in this book, the author would pick one person at random *ex post* and pay a small fraction of earnings, usually several dollars. This procedure is not generally necessary, but it was followed here to reduce unexpected differences between classroom and research data.

Replication

One of the main advantages of experimental analysis is the ability to repeat the same setup numerous times in order to determine average tendencies that are relatively insensitive to individual or group effects. Replication requires that instructions and procedures be carefully documented. It is essential that instructions to subjects be written as a script that is followed in exactly the same manner with each cohort that is brought to the laboratory. Having a set of written instructions helps ensure that unintended “biased” terminology is avoided, and it permits other researchers to replicate the reported results. The general rule is that enough detail should be reported so that someone else could replicate the experiment in a manner that the original author(s) would accept as being valid, even if the results turned out to be different. For example, if the experimenters provide a number of examples of how prices determine payoffs in a market experiment, and if these examples are not contained in the written instructions, the different results in a replication may be due to differences in the way the problem is presented to the subjects.

Control

A second main advantage of experimentation is the ability to control the factors that may be affecting observed behavior, so that extraneous factors are held constant (controlled) as the treatment variable changes. Control can be reduced or lost when procedures make it difficult to determine the incentives that participants actually faced in an experiment. The use of biased terminology may allow participants to act on “homegrown values” that conflict with or override the induced money incentives. For example, experiments pertaining to markets for emissions permits typically do not use the word “pollution.” If people are trading physical objects like university sweatshirts, differences in individual valuations make it hard to reconstruct the nature of demand in a market experiment. There are, of course, situations where non-monetary rewards are desirable, such as experiments designed to test whether ownership of a physical object makes it more desired (“the endowment effect”). Thus control should always be judged in the context of the purpose of the experiment.

Another factor that can disrupt control is the use of deception. If subjects suspect that announced procedures are not being followed, then the announced incentives may not operate in the intended ways. This counter-productive incentive effect is probably the main reason that deceptive practices are much less common in incentive-based economics experiments than is the case in social psychology experiments. Even if deception is hidden successfully during the experiment, subjects may well find out afterwards by sharing their experiences with friends, so the careful adherence to non-deceptive practices provides a

“public good” for those who run experiments in the same lab at a later time. Ironically, the perverse incentive effects of deception in social psychology experiments may be aggravated by an *ex post* confession or “debriefing,” which is sometimes required by human subjects committees.

Psychological Biases

There is a rich behavioral literature in economics and psychology that documents psychological aspects of decision making that may have strong effects on economic behavior. These effects are sometimes called *biases* or *anomalies*, and many of them are summarized in Kahneman, Slovic, and Tversky (1982), Samuelson and Zeckhauser (1988), Tversky and Thaler (1990), Laibson (1997), and in Richard Thaler’s provocative publications (e.g. Thaler, 1988, 1989, 1992). Anomalies and biases will be considered in detail in later chapters, but it is useful to mention some of those that are most relevant for experimental design:

- *Loss Aversion*: This is the tendency for losses to provide a stronger stimuli than gains. It is not uncommon for subjects to react strongly and erratically to losses, and it would be a serious mistake to have losses be more prevalent in one treatment than in another, unless this difference is the focus of the experiment or the necessary result of the treatment.
- *Status Quo Bias*: This is the tendency for subjects to maintain a decision or condition that is established by others or by the experimenter. A closely related idea is the notion of *anchoring*, whereby options are evaluated with reference to an initial position or salient property. Examples used in instructions may suggest focal starting points that can be a problem if they seem to affect results. This can be avoided by using numbers that are quite different from those that will be encountered in the experiment, or by letting subjects provide their own examples and then having the experimenter check calculations for those examples.
- *Endowment Effect*: Often it seems to be the case that ownership of an item or option increases its value to the owner. In particular, the elicitation of a *willingness to accept* sale price by an owner will typically generate higher values than the elicitation of a *willingness to pay* by a prospective buyer. A low willingness-to-pay for a gamble that has randomly determined payoffs indicates risk aversion, but the same person if given the gamble may demand a high selling price or “willingness-to-accept,” which would indicate risk loving behavior. It is essential that results of an experiment not be attributed to biases induced by ownership, unless this is the purpose of the experiment.

Context in Laboratory and Field Experiments

An important design decision for any experiment pertains to the amount and richness of context to provide. A little economic context can be very useful. For example, it is possible to set up a market with a detailed and tedious description of earnings in abstract terminology that does not mention the word “price.” But market terminology helps subjects figure out “which way is up,” i.e. that sellers want high prices and buyers want low prices. Nevertheless, it is an accepted practice in economics experiments to strip away a lot of social context that is not an essential part of the economic theories being tested. If the theories being evaluated do not depend on assumptions about social context, then the best approach is often to try to hold this context constant as the economic parameters are changed. This process of holding context constant may involve minimizing its unintended and unpredictable effects, e.g. by taking steps that increase anonymity during the experiment. Even then, a lot can be learned by re-introducing social context in a controlled manner, e.g. by comparing individual and group decisions.

Social context can sometimes be critically important, as in some politics experiments where it is not possible to recreate the “knock on the door” or phone campaign solicitation in the lab. In such cases, researchers use *field experiments* involving people in their natural environments, who may not even know that they are participating in an experiment. For example, Gerber and Green (2000) targeted political messages to randomly selected voters, using phone, personal contact, or mail, in order to evaluate the effects of these messages on voter turnout, which was determined by looking at precinct records after the election. Field experiments can also provide more relevant groups of subjects and can be used to avoid *experimenter demand effects*, i.e. situations in which behavior in the lab may be influenced by subjects’ perceptions of what the experimenter wants or expects. There are, of course, intermediate situations where the lab setup is taken to the field, to use traders from particular markets in a context that they are familiar with. For example, List and Lucking-Reiley (2000) set up auctions for sports cards at a collector’s convention. Another type of *enriched laboratory experiment* involves making the lab look and feel more like the field situation, as in voting experiments where subjects are seated in a comfortable room decorated like a living room and are shown alternative campaign ads that are interspersed with clips from a local news show.

Although field experiments can introduce a more realistic social context and environment, the cost is often a partial loss of control over incentives, over measurement of behavior, or over the ability to replicate under identical conditions. For example, the effects of alternative political ads on voting behavior in an enriched laboratory setting are typically measured indirectly by surveys of voters’ intentions, since actual votes for one candidate or another are

not public information. In addition, replication may be complicated by interactions between political ads used in the experiment and the positive or negative dynamics of an ongoing political campaign. In other cases, reasonably good controls are available, e.g. even though individual valuations for specific sports cards are not induced directly, they can be approximately controlled by using matched pairs of cards with identical book values.

To the extent that social context and target demographics are important in a field experiment, each field experiment is in some sense like a data point that is specific to that combination of subjects and context unless appropriate random selection of subjects is employed. Thus the results from a series of field experiments become more persuasive, just as do the results from series of laboratory experiments, which is a point made convincingly by Kagel and Roth (1995). There can also be important interactions between the two approaches, as when results from the lab are replicated in the field, or when a general, but noisy, pattern discovered in diverse field situations shows up in a laboratory setting that abstracts away from the diverse field conditions. Parallel laboratory and field experiments will be discussed in some of the chapters that follow.

Independent Observations

In order to reach conclusions supported by standard statistical arguments, it is necessary to have independent observations. For example, suppose you run a market with communication among sellers and get an average price of \$10, and you run it again without communication and get an average price of \$9. Even though this outcome is consistent with your original hypothesis that communication would facilitate price increases, a strong statistical argument cannot be made on the basis of the overall average prices for these two sessions alone (at least not without further statistical modeling of the economic processes within each session). This is because there is always some randomness in prices, and under the null hypothesis of no effect, it is just as likely that the price in the communication session yields a higher price as a lower price. On the other hand, if you ran three separate market sessions with communication and observed higher prices than with three other no-communication sessions, then the chances of seeing this pattern under the null hypothesis are much lower. In particular, think of each communication session as a quarter and each no-communication session as a dime. There are 20 different ways that the quarters and dimes can be arrayed along a horizontal line that represents average prices (question 5), and of these, the most extreme outcome (all three quarters on the high price side) was observed, so the chance is only $1/20$ that this could happen under the null hypothesis in which all 20 outcomes are equally likely. Thus the null hypothesis of no communication effect could be rejected at the $1/20 = 5$ percent level (when the alternative hypothesis is that communication raises prices). (See questions 3-5

for a mathematical formula that can be used to construct these counting arguments.) These arguments do not depend on specific distributions with parameters for the mean, variance, etc. (like the normal distribution), and hence, the resulting tests are called *nonparametric* tests. The clearest presentation of the kinds of nonparametric statistics commonly used by experimenters can be found in Siegel (1956) or Siegel and Castellan (1988), which both contain many examples from economics and psychology. These types of statistical arguments will be encountered in later chapters, but the main idea at this point is that making a statistical claim depends on getting sufficiently strong results with enough independent observations.

Independence of observations can be lost due to contamination. For example, if you had 4 pairs of people negotiating over the division of \$10 and if the first agreement reached were to be announced, then it might affect the remaining 3 agreements. A more subtle case of contamination may with re-matching, so that people are dealing with different partners in the second round. Without announcements, there would be 4 independent bargaining outcomes in round 1, but the round 2 results might depend on subjects' experience in the first round, which could result in contamination and loss of independence. If random matching is desired in order to make each round more like a single-period game, then the standard approach is to treat each group of people who are being re-matched in a series of rounds as a single independent observation. In this case, the experimenter would need to bring in a number of separate groups for each of the treatments (or treatment orders) being investigated. If this seems like an unnecessarily conservative approach, remember that the experimenter is often trying to persuade skeptics about the importance of the results.

Finally, it is important to qualify this discussion by noting that a lot might be learned from even a very small number of market trading sessions, each with many (e.g. hundreds) of participants. The analysis requires careful econometric modeling of the dynamic interactions within a session, so that the un-modeled factors can reasonably be assumed to be independent shocks. Think of it this way, the U.S. macro-economy is a single observation with lots of interactions, but this does not paralyze macroeconomists or invalidate econometric models of macroeconomic phenomena where much of the interest is in the interactive dynamics. But if the process being studied does not require dynamic interactions of large numbers of participants, then a design with more independent observations allows the researcher to reach conclusions without relying on extra modeling assumptions. Replication is one of the main advantages of experimental methods.

Fatal Errors

Professional economists often look to experimental papers for data patterns that support existing theories or that suggest desirable properties of new theories and public policies. Therefore, the researcher needs to be able to distinguish between results that are replicable from those that are artifacts of improper procedures. Even students in experimental sciences should be sensitive to procedural matters so that they can evaluate others' results critically. Moreover, experiments can provide a rich set of topics for papers and senior theses.

Those who are new to experimental methods in economics should be warned that there are some fatal errors that can render the results of economics experiments useless. As the above discussion indicates, these include:

- inadequate or inappropriate incentives,
- non-standardized instructions and procedures,
- inappropriate context,
- uncontrolled effects of psychological biases,
- an insufficient number of independent observations,
- loss of control due to deception or biased terminology,
- the failure to provide a calibrated baseline treatment,
- and the change in more than one design factor at the same time.

IV. A Brief History of Experimental Economics

Figure 1.1 shows the trends in published papers in experimental economics. The first papers by Chamberlin and some of the game theorists at RAND were written in the late 1940's and early 1950's. In addition, early interest in experimental methods was generated by the work of Fouraker and Siegel (1963). (Siegel was a psychologist with high methodological standards; some of his work on "probability matching" will be discussed in a later chapter.) In the late 1950's, business school faculty at places like Carnegie-Mellon became interested in business games, both for teaching and research. And Vernon Smith's early market experiments were published in 1962. Even so, there were less than 10 publications per year before 1965, and less than 30 per year before 1975. Much of the interesting work during this period was being done by Reinhard Selten and other Germans, and there was an international conference on experimental economics held in Germany in 1973. During the late 1970's, Vernon Smith was a visitor at Caltech, where he began working with Charles Plott, who had studied at Virginia under James Buchanan and was interested in public choice issues. Plott's (1979) first voting experiments stimulated work on voting and agendas by political scientists in the early 1980's. Other interesting

work included the Battalio, Green, and Kagel (1981) experiments with rats and pigeons, and Al Roth's early bargaining experiments, e.g. Roth and Malouf (1979). There were still fewer than 50 publications per year in the area before 1985. At that time, my thesis advisor and former colleague, Ed Prescott, told me that "Experimental economics was dead end in the 1960's and it will be dead end in the 1980's."

In the 1980's, Vernon Smith and his colleagues and students at Arizona established the first large laboratory and began the process of developing computerized interfaces for experiments. Arlington Williams (1980) wrote the first posted-offer program. After a series of conferences in Tucson, the Economic Science Association was founded in 1986, and the subsequent presidents constitute a partial list of key contributors (Smith, Plott, Battalio, Hoffman, Holt, Forsythe, Palfrey, Cox, Schotter, Camerer, Fehr, and Kagel).

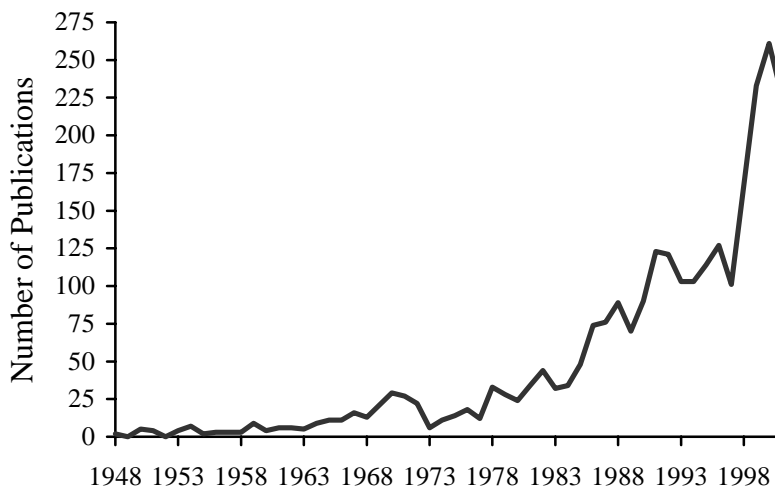


Figure 1.1. Numbers of Published Papers in Experimental Economics

Vernon Smith and Mark Isaac begin editing *Research In Experimental Economics*, a series of collected papers appearing about every other year since 1979. The first comprehensive books in this area were published in the 1990's, e.g. Davis and Holt (1993) and Hey (1994). The 1995 *Handbook of Experimental Economics*, edited by Kagel and Roth, contains survey papers on key topics like auctions, bargaining, public goods, etc. These surveys have some "bite," and they are still remarkably good sources for reference. The first specialty journal, *Experimental Economics*, was started in 1998. The strong interest among Europeans is indicated by the fact that one of the founding coeditors was the

University of Amsterdam, where there is a separate Department of Experimental Economics. The Economic Science Association now has an annual international meeting and additional regional meetings in the U.S. Europe, and Asia. These developments have resulted in over a hundred publications in this area every year since 1990, with highs of over 200 per year since 1999. A searchable bibliography of over 4,000 papers in experimental economics and related social sciences can be found in the author's *Y2K Bibliography of Experimental Economics*: <http://www.people.virginia.edu/~cah2k/y2k.htm>.

There are lots of exciting developments in this field. Economics experiments are being integrated into introductory courses and the workbooks of some major texts. Theorists are looking at laboratory results for applications and tests of their ideas, and policy makers are increasingly willing to look at how proposed mechanisms perform in controlled tests before risking a full-scale implementation. Experimental methods have been used to design large auctions (e.g. the FCC spectrum auctions) and systems for matching people with jobs (e.g. medical residents and hospitals). Two of the recipients of the 1994 Nobel Prize in Economics, Nash and Selten, were game theorists who had run their own experiments. Most significantly, the 2002 Nobel Prize in Economics was awarded to an experimental economist (Smith) and to an experimental psychologist (Kahneman) who is widely cited in the economics literature. Economics is well on its way to becoming an experimental science!

Questions

One of the themes of this book is that “learning by doing” can help students discover and understand key concepts at a deeper level. In addition to participating in experiments, it is important to try out one's understanding by working problems. Some of the questions that follow provide a brief introduction to the way non-parametric statistical arguments are used by experimentalists.

1. It has been observed that collusion on price may be hard to establish if it involves unequal sacrifices, but that once established, collusive agreements in experiments can be stable under some conditions, e.g. if mid-period changes in posted prices are not allowed. In an experiment to evaluate the effects of communication opportunities on price levels, would you prefer to have each group of sellers subjected to a single treatment, or would you prefer to have each group experience some periods with communication and some without, perhaps alternating the order? What considerations might be relevant?
2. A consulting firm conducted a random survey of residents of a community, describing a planned riverside park and then asking each respondent the question: “what is the most that you would be willing to

pay to have this park built along the river?” Do you think that the monetary benefits estimated from the responses are likely to be an overestimate or an underestimate? What is the source of the bias?

3. Consider a between-subjects design with two treatments: “dime” and “quarter.” If there are just two separate sessions, one with each treatment, and if the one with the lowest price outcome is listed on the left, then the two possible ranking outcomes can be represented as DQ and QD. Suppose instead that there are two sessions with each treatment, so that one of the possible rankings is DDQQ. Find the other five rankings.
4. The setup in the previous problem, with two sessions in each of two treatments, yields 6 possible rankings of the D and Q designators. Mathematically, this represents the number of ways of ranking 4 things by type, when there are two things of each type. The mathematical formula for this is: $(4 \cdot 3 \cdot 2 \cdot 1)$ divided by $(2 \cdot 1) \cdot (2 \cdot 1)$, or $(4!)/(2! \cdot 2!)$, which yields $24/4 = 6$. Now consider a setup with a total of six sessions, with three sessions for each of two treatments, D and Q. Either calculate or count the number of rankings by type, and explain your answer. (It is not permissible to simply quote the number given in the chapter.)
5. Think about the intuition behind the ratio of factorial expressions in the previous question. With 4 items, there are 4 ways to pick the first element in a sequence, 3 ways to pick from the remaining 3 elements, 2 ways to pick the third element, and only 1 element left for the final position, so the number of possible orders is $4 \cdot 3 \cdot 2 \cdot 1$. Thus the numerator of the formula $(4!)$ is the total number of possible rankings for the sessions. Division by the factorial expressions in the denominator reduces the number in the numerator. This is done because all that is required in the statistical argument is that the sessions be identified by the treatment, e.g. Q or D, and not by the particular session done with that treatment. For example, if Q_1 is the first session done with the Q treatment, and Q_2 is the second session, then the two orderings, Q_1Q_2 and Q_2Q_1 , are only counted once, since the theory does not make a prediction about order within the same treatment. Using this observation, explain why the denominator in the formula is the product of two factorial expressions.
6. Consider a “within-subjects” design in which each person (or group) is exposed to two treatments, e.g. two numerical decisions (“D1” and “D2”) are made under differing conditions, with one decision to be chosen at random *ex post* to determine earnings. If $D1 > D2$, then code this as a Heads for that person. A natural null hypothesis would be that the probability of Heads is $1/2$, so the chances of two Heads in two trials is $(1/2)(1/2) = 1/4$. If there are 5 subjects and Heads is observed in all cases, what are the chances that this could have occurred at random?