1

Designing and Conducting Laboratory Experiments

Elena Katok

Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX, USA

1.1 Why Use Laboratory Experiments?

Operations management (OM) is a field with strong tradition of analytical modeling. Most of the early analytical work in OM was primarily optimization based and dealt with central planning for such problems as job-shop scheduling, lot sizing, and queuing. Starting in the 1980s, OM researchers became interested in modeling strategic settings that involve interactions between firms. Today, OM models tackle problems that deal with supply chain coordination, competition, and cooperation, which examine incentives and objectives of firms as well as individual decision makers. This type of work requires a model of decision-making at individual and/or firm level.

Supply chains are not centralized, but consist of individual self-interested firms – original equipment manufacturers (OEMs), different tiers of suppliers, transportation vendors, and retailers. These firms face uncertainty from the environment, such as production yield, processing times, and customer demand, as well as strategic uncertainty, which comes from the uncertainty about the actions of the other supply chain members. Traditionally, OM models assumed that firms are expected profit maximizers and are fully rational, meaning that they correctly anticipate the actions of the other supply chain members.

Behavioral operations management (BOM) started in order to first test, and then improve, modeling assumptions about decision-making. Schweitzer and Cachon (2000) is the seminal BOM paper that tested how individuals solve the "newsvendor problem." It turned out that individuals generally do not solve the problem correctly, but are rather systematic and predictable in how their

The Handbook of Behavioral Operations, First Edition. Edited by Karen Donohue, Elena Katok, and Stephen Leider.

decisions deviate from optimal. Schweitzer and Cachon (2000) finding, and numerous studies that followed (see Chapter 11), has major implications for OM models, because the newsvendor problem is a building block for much of the inventory theory.

BOM work lives at the boundary of analytical and behavioral disciplines. It is aimed at developing models of decision-making to better explain, predict, and improve analytical models in OM. There are many empirical methods for studying human behavior in general and human judgment and decision-making in particular. Laboratory experiment, the topic of this chapter, is one of the empirical methods we use in BOM. Similar methods have been employed in a number of other social science fields, including psychology and sociology (social networks), law (jury behavior), political science (coalition formation), anthropology, biology (reciprocity), and especially experimental economics, that have a long and rich tradition of studying problems that are similar to the ones of interest to the OM community.

Laboratory experiments can be designed to test analytical models in a way that gives the theory the best possible shot to work. This is done by carefully controlling the environment, especially information available to the participants, to match theoretical assumptions. Parameters can be selected in a way that treatment effects predicted by the model are large enough to be detected in the laboratory, given appropriate sample sizes and the level of general "noise" in human behavior. If the theory fails to survive such a test, a conclusion can be made that the model is likely to be missing some important behavioral aspect. If a theory survives such a test, we can conclude that that the model qualitatively captures enough of the behavioral factors to organize the data, and further robustness tests can be performed by manipulating parameters.

The ability to cleanly establish causality is a relative advantage of laboratory experiments, compared with other empirical methods. In the laboratory, causality is established by directly manipulating treatment variables at desired levels and randomly assigning participants to treatments. Random assignment ensures that treatment effects can be attributed to the treatment variables and not be confounded by any other, possibly unobservable, variables. Other empirical methods rely on existing field data, so neither random assignment nor direct manipulation of treatment conditions is possible, so causality cannot be directly established.

Another advantage of laboratory experiments is that they lend themselves well to being replicated by researchers in different laboratories. Replicating results is important because any single laboratory result can be an artifact of the protocols or settings in the specific laboratory.

Results that have been replicated in different contexts and by different research teams can be considered reliable. A recent article published in *Science* (Open Science Collaboration 2015) highlighted the importance of replicating

experimental results. It reported that only 36% of psychology studies published in three important psychology journals and selected as part of a large-scale replication project had statistically significant results when replicated. Replications done in the *Science* article showed that while in the original studies most reported results were large in magnitude and statistically significant, in replications, most results were smaller in magnitude and not significant, although mostly directionally consistent with the original results. A similar study of economics experiments (Camerer et al. 2016) reports that approximately 2/3 of the economics studies replicated successfully.

Laboratory studies complement other methods by bridging the gap between analytical models and real business problems. Analytical models are built to be parsimonious and general and are primarily normative in nature. They use assumptions to make the mathematics tractable. These models can be tested using a variety of empirical methods, including surveys, field studies, field experiments, or laboratory experiments. Empirical methods are, by their nature, descriptive. All empirical methods involve a trade-off between the internal and the external validity. Surveys and field studies that use secondary data have high external validity (they are close to the real settings being studied), but may be low on internal validity (the ability to establish the cause-and-effect relationship based on the data) because they often suffer from being confounded, or not having all the data that would ideally be required. This is because researchers cannot directly manipulate the factors or levels in the study – they have to accept data that is available to them. Experiments need no take place in the laboratory – they can take place in the field also. Field and lab experiments usually differ in their level of control and in their ability to establish causality. Laboratory experiments are high on the internal validity, but because the environment is often more artificial, they are lower on the external validity.

1.2 Categories of Experiments

According to Roth (1995a), laboratory experiments fall into three broad categories. The first is to *test and refine existing theory*. Much of the BOM work so far fell into this category. For example, experiments testing behavior in the newsvendor model (Schweitzer and Cachon 2000; Bolton and Katok 2008) test how well people are able to optimize under uncertainty. The second category has to do with *characterizing new phenomena leading to new theory* that help organize behavioral regularities. An example is the literature on social preferences. In the OM domain, Loch and Wu (2008) found in a lab experiment that concerns with status and relationship have an effect on the performance of the wholesale price contract. Cui, Raju, and Zhang (2007) develop a fairness model and apply it to the setting of a wholesale price contract to formally characterize conditions that may lead to channel coordination with the wholesale pricing.

Özer, Zheng, and Chen (2011) found that people are more truthful than standard theory suggests and develop a model of trust and trustworthiness that explains some of the regularities in their lab experiment. The third category deals with *testing institutional designs*. Some institutions are not well understood, and the laboratory can be used to gain insights in their performance. There are several notable examples in economics, such as designing the Federal Communications Commission (FCC) auctions for radio spectrum (Goeree and Halt 2009) or designing the market for medical interns (Roth 1984).

A good experiment is one that controls for the most plausible alternative hypotheses that might explain the data. It also allows the researcher to cleanly distinguish among possible explanations. For example, the Schweitzer and Cachon (2000) study looks at the behavior in the newsvendor problem. In the setting in which the critical fractal is above 0.5 (called the high profit condition), the authors find that average orders are below the optimal order and above the mean demand. At this point a potential plausible explanation is risk aversion – risk-averse newsvendor should order less than the risk-neutral newsvendor. But the Schweitzer and Cachon (2000) design cleverly includes a low profit condition, with the critical fractal below 0.5. In that treatment, risk aversion still implies that orders should be below optimal, but the authors find that orders are above optimal. Thus, the design can clearly rule out risk aversion as the (only) explanation.

Three factors make experimental work rigorous. The first one is *theoretical guidance*. To interpret the results of an experiment, researchers need to be able to compare the data to theoretical benchmarks. Systematic deviations from theory can provide insights into factors missing from the analytical model, and guidance into how the model can be improved.

The second factor is *induced valuation*. In his seminal paper, Smith (1976) explains how a reward medium (for example, money) can be used to control the objectives of the laboratory participants. When participants are rewarded based on their performance in the experiment, researchers have a cleaner test of how people pursue their goals. This test is not confounded by not knowing what those goals are.

The third factor is careful *control of institutional structure*. Strategic options and information available to participants should match those assumed by the theoretical model. For example, real bargaining is typically done face-to-face and is often unstructured, making modeling bargaining extremely challenging. But some assumptions can be imposed on the bargaining process to make a model tractable while still capturing some essential features of real bargaining. For example, we may assume that bargainers exchange alternating offers, and to capture the fact that no bargaining process can go on forever, we may assume that the pie they are bargaining over is discounted at each iteration. These two assumptions allow for a tractable model (Rubinstein 1982 that provides useful insights and has clear empirical predictions. A model can be further

streamlined by assuming that the bargaining process is finite. It turns out that what the model predicts about how the pie will be split depends on length of the bargaining process and the relative discount rates of the two players. These predictions cannot be tested in the field because real bargaining processes are substantially different from the model, but the model can be tested in the laboratory. To continue with another example from the bargaining literature, Ochs and Roth (1989) found that in a two-period version of this bargaining game, players in the second period often make offers that are less in absolute terms than the original first period offers they received. These "disadvantageous counteroffers," however, are better in relative terms. Bolton (1991) showed, among other things, that these fairness concerns are significantly reduced when players are paid based on a tournament structure. The results of these, and many other tests, provided seminal insights that formed the basis for the theory of social preferences (Fehr and Schmidt 1999; Bolton and Ockenfels 2000).

One of the questions that are often asked about laboratory experiments is whether their results can be carried over into the real world. Smith (1982) addresses this question with the concept of *parallelism*. He writes: "Propositions about the behavior of individuals and the performance of institutions that have been tested in laboratory micro economies apply also to non-laboratory micro economies where similar *ceteris paribus* conditions hold" (Smith 1982, p. 936). In other words, behavioral regularities persist as long as relevant underlying conditions are substantially unchanged.

The art of designing good experiments (as well as the art of building good analytical models) is in creating simple environments that capture the essence of the real problem while abstracting away all unnecessary details. Thus, the first step in doing experimental work is to start with an interesting theory. What makes a theory interesting is that (i) it has empirical implications and (ii) these implications are worth testing, meaning that they capture a phenomenon that is sufficiently real and interesting so that learning about it adds to our knowledge of the real world.

This chapter focuses on controlled laboratory experiments used to test existing, and develop new, theory in OM. Much of the methodology I discuss is in line with economics rather than psychology, which also provide a valid and useful, but different, paradigm. The rest of this chapter is organized as follows: In Section 1.2 I discuss some fundamental games that proved to be important in economics as well as in BOM. These games will come up again in several other chapters in this book. In Section 1.3 I will discuss some basics of experimental design as well as "best practices" for conducting laboratory experiments. In that section I will touch on issues related to providing a context, the effect of subject pool, the effect of incentives, and the uses of deception. I conclude this chapter with a discussion of my view of future trends and promising directions for future research.

1.3 Some Prototypical Games

1.3.1 Individual Decisions

The desire to test whether people behave consistently with mathematical models is perhaps as old as the desire to analytically model human behavior. This literature is the subject of Chapters 3 and 5 in this handbook. The well-known St. Petersburg paradox (Bernoulli 1728) was the first to illustrate the problem with modeling people as maximizing their expected profits. It goes as follows: A fair coin is tossed until it comes up heads. You get \$1 when it lands on heads the first time, \$2 when it lands on heads the second time, \$4 when it takes three tosses, and \$8 when it takes four tosses. Name the greatest certain amount that you would pay to play this game once. The expected value of this bet is $\sum_{i=1}^{\infty} i(1/2)^i$ and does not converge. Yet most people would value this lottery at about \$20. Bernoulli proposed a "utility function" with diminishing marginal utility so that the sums converge.

There were early experiments on individual choice testing ordinal utility theory, starting as early as Thurstone (1931), who estimated individual's indifference curves through a large sequence of hypothetical questions. Almost immediately, and as a reaction to this work, Wallis and Friedman (1942) criticized it for basing the analysis on hypothetical choices and encouraged future experiments in which subjects are confronted with real, rather than hypothetical, choices.

After the publication of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (Von Neumann and Morgenstern 1944), various aspects of expected utility theory were tested; the most famous of those tests is known as the *Allais paradox* (Allais 1953). Allais presented his subjects with two hypothetical choices. The first was between alternatives A and B:

A: 100 million francs with certainty
B: 10% chance of 500 million francs
89% chance of 100 million francs
1% chance of 0

The second was between alternatives C and D:

C: 11% chance of 100 million francs
89% chance of 0
D: 10% chance of 500 million francs 90% chance of 0

An expected utility maximizer who prefers A to B should also prefer C to D, but a common pattern observed was to prefer A to B and D to C. This experiment has been subsequently replicated using (much smaller) real stakes.

The Allais paradox is only one of many violations of the expected utility theory, and identifying numerous other violations and modifying or extending the model to account for these violations produced an enormous amount of literature at the intersection of economics and cognitive psychology. See Machina (1997) for an overview and Camerer (1995) for a detailed literature survey of individual decision-making, as well as Chapter 5 of this handbook.

In spite of numerous documented violations, the expected utility theory continues to be the predominant paradigm in economics as well as in the OM. One reason for this is that although numerous alternatives have been proposed, none are as elegant or analytically tractable as the original model. Thus, in OM, in spite of Bernoulli's early demonstration in 1728, the majority of models assume expected profit maximization, and even allowing for risk aversion is a fairly new phenomenon.

1.3.2 Simple Strategic Games

Following von Von Neumann and Morgenstern (1944), economists also became interested in testing models of strategic interactions. Chapter 7 of this handbook provides a detailed review of this literature. One of the first strategic games studied in the laboratory is known as the *prisoner's dilemma* (Flood 1958). In this game two players (labeled Row and Column) must simultaneously choose one of two options (that for transparency we will label Cooperate and Defect, but that carried neutral labels "1" and "2" in the experiments). The payoffs are displayed in Figure 1.1.

Both players in the prisoner's dilemma game have the *dominant strategy*. A player has a dominant strategy when her preferred option does not depend on the choice of the other player. Observe that the Column player earns more from defecting than from cooperating regardless of what the Row player does (2 vs. 1 if Row cooperates and ½ vs. –1 if Row defects). Similarly, the Row player earns more from defecting than from cooperating regardless of what the Column player does (1 vs. ½ if Column cooperates and 0 vs. –1 if Column

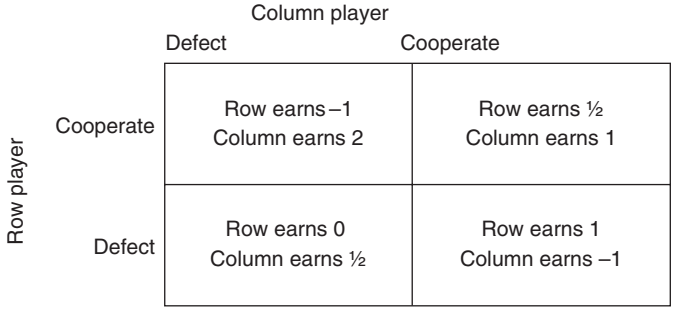

Figure 1.1 Payoffs in the prisoner's dilemma game (Flood 1958).

defects). Thus, the unique equilibrium in the prisoner's dilemma game is for both players to defect, Row earning 0 and Column earning ½. This outcome is inefficient, because both players can be better off from cooperation.

Players in the Flood (1958) study played 100 times, and average earnings were 0.4 for Row and 0.65 for Column – far from the equilibrium prediction but also far from perfect cooperation. The authors interpreted their results as evidence against the equilibrium solution, but also included in their paper a comment by John Nash, who pointed out that in a game repeated 100 times, while Defect continues to be the unique equilibrium, other strategies are also nearly in equilibrium,[1] so the experiment to test the theory should be conducted with *random matching* of the players. The game of prisoner's dilemma continued to fascinate social scientists for decades, and still does, because of its broad applications. It has been "…used as a metaphor for problems from arms races to the provision of public goods" (Roth 1995a, p. 10).

Another topic deeply rooted in experimental economics that has important implications for OM is bargaining. Güth, Schmittberger, and Schwarz (1982) were the first to conduct an experiment on the ultimatum game, which has since become the standard vehicle for modeling the negotiation process. The game involves two players. The Proposer received $10 and has to suggest a way to distribute this amount between himself and the other player, the Recipient. The Recipient, upon observing the Proposer's split, can either accept it, in which case both players earn their respective amounts, or reject it, in which case both players earn 0. The ultimatum game has the unique subgame perfect equilibrium that can be computed using backward induction. Looking at the responder's decision first and assuming the responder would prefer any positive amount of money to 0, it follows that the responder should be willing to accept the smallest allowable amount (1 cent). Knowing this, the responder should offer 1 cent to the responder and take $9.99 for himself. In fact, Proposers offer a split that is closer to 60% for themselves and 40% for the responder, and moreover, responders tend to reject small offers.

Since the Güth, Schmittberger, and Schwarz (1982) experiments were conducted, hundreds of ultimatum experiments have been reported. Roth et al. (1991) conducted a large-scale study in four countries: the United States, Yugoslavia, Japan, and Israel. In each country they compared the ultimatum game (one proposer and one responder, called "Buyer" and "Seller") and the market game (one "Seller" and nine "Buyers"). In the market game the buyers submit sealed bids, and the seller can accept or reject the highest offer. They found that in all four countries, the market game quickly converged to the

1 For example, in the "tit-for-tat" strategy, players start by cooperating and then mimic the behavior of the other player in the previous round (Alexrod 1984).

equilibrium prediction, in which the seller receives nearly the entire pie, while the results of the ultimatum game showed no signs of converging to this equilibrium. There were some differences reported in the ultimatum game among the four countries.

Ochs and Roth (1989) report on a series of two-stage bargaining experiments in which player 1 makes an offer, player 2 can accept or reject, and if player 2 rejects, the pie is discounted (multiplied by $\delta < 1$), and player 2 can make an offer to player 1. Player 1 can then accept or reject, and if player 1 rejects, both players earn 0. We can work out the equilibrium again using backward induction. Starting with stage 2, player 2 should be able to earn the entire discounted pie, which is δ. Knowing this, player 1 should offer player 2 δ in the first stage, and player 2 should accept it.

Ochs and Roth (1989) report two regularities:

1) Disadvantageous counteroffers: Player 2 in the second stage makes an offer that gives himself (player 2) less than player 1's offer in stage 1.
2) The deadline effect: Most agreements happen in the last second.

In regard to the disadvantageous counteroffers, Ochs and Roth (1989) conclude: "We do not conclude that players 'try to be fair.' It is enough to suppose that they try to estimate the utilities of the player they are bargaining with, and […] at least some agents incorporate distributional considerations in their utility functions" (p. 379).

Forsythe, Horowitz, and Sefton (1994) specifically explore the question of what motivates proposers in the ultimatum game. To do this, they conducted the *dictator game*. The dictator game is almost the same as the ultimatum game, but the responder does not have the right to veto an offer. This means that there are no strategic reasons to yield any ground. Contributions reflect "pure" preferences. I will discuss the Forsythe, Horowitz, and Sefton (1994) paper in more detail in the following section as a way to illustrate the importance of using monetary incentives. I refer the reader to Roth (1995b) for a review of bargaining experiment prior to 1995. This literature also gave rise to both analytical and behavioral literature on other-regarding preferences (that is, incorporating concerns for others' earnings directly into the utility function). This material is covered in Chapter 6 of this handbook. Cooper and Kagel (2016) provide a review of post-1995 economics literature on social preferences. Chapter 13 of this handbook covers related BOM literature on supply chain contracting.

1.3.3 Games Involving Competition: Markets and Auctions

A central pillar of economic theory is the principle that prices clear markets. Competitive equilibrium (CE) prices are determined at a point at which supply meets demand, but how exactly prices arrive at this level is (still) not well

understood. Adam Smith famously termed this the "invisible hand." Some practical questions that economists disagreed on regarding the requirements for CE prices to come about included the number of buyers and sellers and the amount of information.

Chamberlin (1948) set out to gain initial insights into this question with a laboratory experiment that involved a large numbers of students in the roles of buyers and sellers. Each buyer had a privately known value, each seller had a privately known cost, and they interacted through a series of unstructured bilateral negotiations. So this market had a large number of traders, but no centralized information. Chamberlin (1948) reported that prices were quite dispersed and showed no tendency of quickly converging to equilibrium, and as a result there was substantial inefficiency.

Smith (1962) conducted a famous experiment in which he essentially repeated Chamberlin's experiment, but added a *double auction* institution that allowed buyers and sellers to make and accept public bids and asks.[2] Additionally, Smith (1962) repeated the market several times, allowing buyers and sellers to keep their costs and valuations for several rounds. The price converged to the equilibrium level reliably and quickly (but not in the first round). Smith's early work on the double auction institution is foundational and generated a long and fertile literature (see Holt 1995).

The behavior of two-sided markets (multiple buyers and multiple sellers) is more complicated than behavior of one-sided markets. Markets with a single seller and multiple buyers are called *forward auctions*, and markets with a single buyer and multiple sellers are called *reverse auctions*.

The field of auction theory is extensive (see Krishna (2002) for a comprehensive review of theoretical literature), and laboratory experiments have been used to test many of these models. I refer the readers to Kagel (1995) for a comprehensive review of work done prior to 1995 and to and Kagel and Levin (2016) for work done since 1995, while Chapter 15 of this handbook focuses on reverse auctions.

1.4 Established Good Practices for Conducting BOM Laboratory

In this section I discuss several methodological topics related to good practices in designing and conducting laboratory experiments.

2 The story is that Vernon Smith initially became interested in this question after he was a subject in Chamberlin's experiment at Harvard (Friedman and Sunder 1994).

1.4.1 Effective Experimental Design

In laboratory experiments, researchers generate their own data, and this, as I already pointed out, allows for better control than in studies that rely on data that occurs naturally. The topic of experimental design is one that deserves a significantly more comprehensive treatment than what I can provide in a short review article. I refer the readers to List, Sadoff, and Wagner (2010) for a brief review and to Atkinson and Donev (1984) for a more detailed treatment, while Fisher (1935) provides a very early textbook on the subject.

When we design an experiment, we are specifically interested in the effect of certain variables, called *focus variables*, but not in the effect of some other variables, called *nuisance variable*. For example, if we are interested in testing a new auction mechanism, we may be specifically interested in the effect of the number of bidders, or the amount and type of feedback – those are focus variables. We may not be specifically interested in the effect of the bidder's experience, or gender, or major – these are nuisance variables. Focus variables should be systematically manipulated between treatments. For example, we may run some treatments with 2 bidders, and some treatments with 4 bidders, to establish the effect of the number of bidders. We call this varying the focus variables at several number of *levels*. In contrast, nuisance variables should be held *constant* across treatments, so that any treatment effects cannot be attributed to the nuisance variables, or to the *interaction effect* between the focus and the nuisance variables. For example, it would be a very poor design to have 2-bidder auctions include only females and all 4-bidder auctions to include all males, because not holding gender constant introduces a confounding interaction effect between the gender and the number of bidders.

The simplest way to avoid inadvertently confounding the experimental design with nuisance variables is to randomly assign participants to treatments from a set of participants recruited from the same subject pool. Thus, it is not advisable, for example, to recruit participants from classes, because doing this may inadvertently assign all subjects from the same class to a single treatment. Similarly, it is not advisable to recruit subjects directly through student organizations, clubs, or fraternities. The idea is to avoid any systematic composition of subjects in a specific treatment.

A good experiment requires at least two treatments, one being the baseline treatment and the second being a comparison treatment. An experiment with only one treatment is not so much an experiment, as it is a demonstration. Sometimes demonstrations can be quite influential and informative (for example, Sterman (1989) is a one-treatment experiment, which is a demonstration of the "bullwhip" effect).

The most straightforward way to construct treatments in an experiment is to simply vary each focus variable at some number of levels and conduct a separate treatment for each combination. This is known as a *full factorial design*.

		Number of bidders	
		$n = 2$	$n = 4$
Auction format	Open bid	OB-2	OB-4
	Sealed bid	SB-2	SB-4

Figure 1.2 An example of a 2×2 full factorial design.

An example of a full factorial design in an experiment with focal variables being the number of bidders and the auction format may be to vary the number of bidders at $n = 2$ or 4 and the auction format at sealed bid or open bid. So the resulting 2×2 full factorial design is shown in Figure 1.2.

The advantage of the full factorial design is that it provides the cleanest evidence for the effect of each variable, as well as all possible interaction effect. But the disadvantage is that in an experiment with a large number of focal variables, a full factorial design can become prohibitively expensive because of the number of subjects required.

A practical way to deal with budget constraints is to use a fractional factorial design instead of full. For example, suppose you have three focal variables and you would like to vary each at two levels, which we denote as A and B. This yields a $2 \times 2 \times 2$ full factorial design with the following eight treatments:

AAA AAB ABA ABB BAA BAB BBA BBB

Suppose you can only afford to run four treatments. The question is, which four to run? Imposing a constraint that the third factor is the product of the first two results in a balanced design (this example can be found in Friedman and Sunder (1994)):

AAA ABB BAB BBA

Another way to construct an experiment when a full factorial design is not feasible is to design treatments in a way that allows you to make a direct comparison with the baseline. This is advisable when you are primarily interested in the effect of individual focal variables, rather than in the interaction effects. For example, the experiment in Katok and Siemsen (2011) uses this design because the experiment contains four focal variables (so the full factorial design would have required 16 treatments, if each was to be varied at two levels). Instead, the authors conducted five treatments:

AAAA BAAA ABAA AABA AAAB

The study investigated the effect of each of the four variables and compares them to the baseline (AAAA) one at a time.

Some nuisance variables cannot be directly controlled (for example, subject's alertness). If you have reason to suspect that there may be some nuisance variable present, you can try to eliminate its effect by *randomizing*. For example, if you believe that subjects who arrive to the lab earlier are better organized and are likely to be more alert, you may try to randomize roles as subject arrive.

A *random block* design holds one or more nuisance variables constant across treatments. An example is a within-subject design that has the same subject participate in more than one treatment. In theory it controls for all possible individual differences among subjects since each subject is exposed to each treatment. In practice, however, within-subject design introduces potential order effect: The order in which treatments are presented to subjects may matter. One method to deal with the order effect is to randomize the order and then statistically test for the order effect. This may not be ideal, however, if the number of treatments is large because failure to detect order effects does not provide a convincing evidence that they are not there, but only that the design does not have sufficient power to detect them.

A clever way to use within-subject design but avoid the order effect is called the *dual trial design*. Kagel and Levin (1986) used this design when they investigated the effect of the number of bidders in a group on bidding behavior in sealed-bid common value auctions. Each decision involved an individual, who, upon seeing his private signal, placed two bids – one for the small group and one for the large group. Both decisions were made on the same screen, so order effects were not an issue. At the same time, the design controlled for all individual differences, so differences in behavior could be fully attributed to the number of bidders.

1.4.2 Context

I will begin with some thoughts on the pros and cons of providing context in experiments. In experimental economics, researchers often describe the experimental tasks to participants using an abstract frame. An abstract frame uses neutral labels for roles and actions. For example, rather than being called "Supplier" and "Buyer," players might be labeled "Player 1" and "Player 2," while possible choices might be described in terms of selecting from a set of options, rather than making business decisions, such as selecting prices and quantities.

There are two reasons for using an abstract frame. One reason is to avoid leading the participants by unintentionally (or intentionally) biased decisions. For example, in an experiment that deals with trust, a participant may have to decide whether to reveal some information truthfully or not. Labeling these actions using loaded language such as "Tell the Truth" or "Deceive" is likely to result in different behavior than labeling the actions "Option A" and "Option B." While the above example is quite stark, often what might be considered leading is in the eye of the beholder. One researcher may think that the language is

neutral, while another researcher (or a referee) may think it is leading. For this reason, using abstract and neutral language is a good practice.

The second reason has to do with a perception that abstract and neutral language somehow makes the experiment more general. If participants are given a specific "cover story," the results are more related to this specific context than to a different context that the same basic setting may also represent. So one school of thought is that because an abstract frame is equally applicable to different settings, the abstract frame is better.

An alternative way to view an abstract frame, however, is that it is not related to *any* real setting. So rather than being more general, it may be less general, because it applies only to a strange and abstract game and not to any business situation to which participants can relate. This point brings us to the main downside of using an abstract frame – it makes the experiment more difficult to explain to participants and may result in more confusion, slower learning, and potentially noisier data.

Unfortunately, there is no simple rule of thumb about context, because one thing is certain: Context matters a great deal. More generally, there is much evidence that *framing* (how the problem is described to participants) can have a large effect on behavior (Tversky and Kahneman 1981). In BOM, we tend to use a cover story that is related to the application we are investigating. This is often reasonable because it may increase the external validity of the experiment and link it closer to the real operations setting under investigation. Researchers should take great care, however, in balancing the need for context with unintentional framing and leading.

1.4.3 Subject Pool

Perhaps one of the first questions people ask about laboratory experiments has to do with the subject pool effect. After all, managers solve business problems; so how valid are results of experiments that use students (mostly undergraduates) as subjects? The first point that is important to emphasize is that laboratory experiments can be conducted with any subject pool. Using students is convenient, but it is not an inherent part of the laboratory methodology. The second point to emphasize is that to the extent that there is any systematic evidence that managers perform any better (or any worse, for that matter) than students, the differences tend to be observed for very specialized set of tasks, and these are typically not the tasks that participants are asked to perform in controlled laboratory experiments. Fréchette (2012) presents a thorough review of experimental results that use nonstudent subject pools and reports that for tasks that have been studied with different subject pools, qualitative results are quite consistent.

There are some obvious practical reasons for using students in experiments. Students are readily available on college campuses, so they can be easily recruited to participate in studies. The cost of providing students with

sufficient financial incentives to take the study seriously and pay attention is relatively low (for planning purposes I use a figure of $20 per hour). It is convenient to invite students to physically come to the lab and participate in a study. This procedure makes it easier to make sure that participants do not communicate, and it is also easier, in this setting, to ensure that all participants have common information. In other words, using students increases the amount of control. It also makes a study easier to replicate.

In my opinion, there are two downsides of using professional managers in experiments. One downside is that it is impractical to incentivize them with money. So either the cost of the experiment rises dramatically, or managers are not directly incentivized with money. Depending on the study, having monetary incentives may or may not be critical – I will discuss the importance of incentives in the next section – but the decrease in control that comes from not having incentive compatibility (having the earnings of the participants be directly related to their actions) should be weighed against the possible benefits of having nonstudent subject pool. Another downside has to do with replicability. An advantage of laboratory experiments is that they can be easily replicated by different research teams. The ability to replicate experiments improves our confidence in reported results and also significantly reduces scientific fraud. A nonstudent subject pool may make a study impossible to replicate.

Does subject pool make a difference? It is quite clear at this point that there is no evidence that managers perform systematically better or worse than students in relatively simple games that we conduct in the laboratory. There are not many studies that systematically considered the subject pool effect; most studies that deal with subject pool do so opportunistically. For example, Katok, Thomas, and Davis (2008) conducted a set of experiments that examine the effect of time horizons on the performance of service-level agreements. They replicated two of the most important treatments in their study with managers (students in an executive education class) who were not incentivized with money, but simply were asked to play the game in order to help the researchers with their study. They report that the only difference between the students' and the managers' behavior is that there is more variability in the manager data than there is in the student data.

Moritz, Hill, and Donohue (2013) investigate the correlation between cognitive reflection test (CRT) scores and the quality of decisions in the newsvendor problem. They have data for students and managers for one of the treatments in their study, and for that treatment the two subject pools perform qualitatively the same. There are also a few other studies that report no difference between the performance of students and professionals in laboratory experiments (Plott 1987; Ball and Cech 1996).

One study that does systematically look at the differences between students and managers is Bolton, Ockenfels, and Thonemann (2012). In the context of the newsvendor game (see Chapter 11 for a review of this literature), the

authors compare performance of three subject pools: undergraduate students (called juniors), masters-level students (called seniors), and managers in an executive education class (called managers). In the experiment, subjects made a sequence of newsvendor decision, and additional information was revealed to them sequentially.

Everyone started knowing the price and cost information that they need in order to compute the critical ratio and were given historical demand information. After 40 rounds (called Phase 1), participants were told that the demand distribution is uniform from 1 to 100. After another 40 rounds (called Phase 2), participants received a tutorial on how to compute the optimal solution and made the last 20 decisions (called Phase 3).

Figure 1.3 summarizes mean order quantities in the Bolton, Ockenfels, and Thonemann (2012) study. All three groups exhibit the pull-to-center effect and do not exhibit any learning within each phase, and all three groups perform better after the tutorial on computing the optimal order quantity (Phase 3). There is no evidence that managers perform any better than the other two groups, and in fact, managers perform slightly worse in Phase 3 than masters students. This study is notable because the experiment is extremely carefully done. The subject pool is the only difference between the treatments – everything else, including the user interface, the instructions, and the incentives, was kept identical.

Managers in the study were procurement professionals, and the analysis in the paper controls for their position in the organization, their education, and their years of experience. While there are some intriguing findings related to

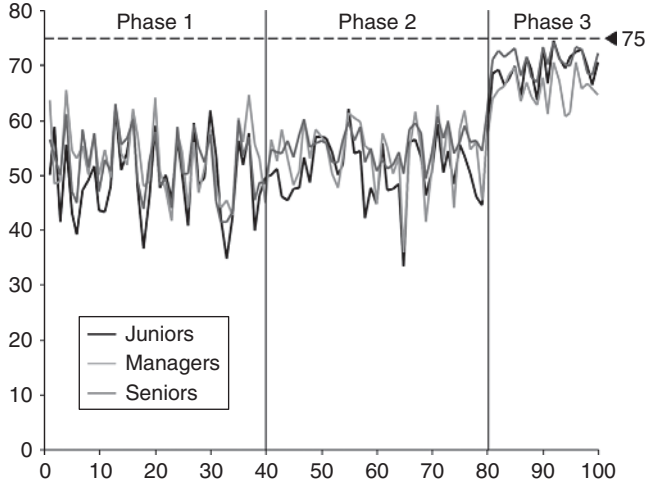

Figure 1.3 Mean order quantities in the Bolton, Ockenfels, and Thonemann (2012) experiment.

this demographic information (higher-level executives tend to do better, for example), the overall result is that managers do not perform better than students do. This result is a typical one related to the subject pool effect. There is no systematic evidence that student subject pool yields different results than professionals. Therefore, using student participants in laboratory experiments is a good procedure and is a reasonable first step, unless there are some very specific reasons to believe that professionals are likely to behave significantly and systematically different.

Mechanical Turk is another source of participants for experimental research that has become popular in a variety of scientific disciplines, including economics, management, and political science. Paolacci, Chandler, and Ipeirotis (2010) investigated several aspects of data quality from samples collected using Mechanical Turk. They report that 70–80% of Mechanical Turk workers are from the United States and participate in Mechanical Turk tasks as a form of entertainment. Many workers generate a small amount of money with this work – on the order of $20 per week, but over 60% report that earning additional money is a major reason for participating. The number of "professional" turkers is rather small (about 14%), but there are some who earn as much as $1000 per month. Overall it seems that Mechanical Turk workers are reasonably well motivated with money, similarly to college students. Mechanical Turk workers are on average more educated than the overall US population and have somewhat lower income than the average US Internet users (with 2/3 earning less than $60K annually). Paolacci, Chandler, and Ipeirotis (2010) found that in terms of age, Mechanical Turk subject pool is significantly older than student subject pool (with the median age of 29 vs. 19 at a large Midwestern university) and somewhat younger than the average US population. The proportion of females is about 2/3, so women are overrepresented.

In a recent paper, Lee, Seo, and Siemsen (2017) used Mechanical Turk to replicate three high BOM studies, originally conducted in the laboratory. They found that social preferences are significantly weaker on the Mechanical Turk than they are in the laboratory. They also found that learning appears to be significantly slower. Nevertheless, most of the qualitative aspects of the data were replicated successfully.

So overall, Mechanical Turk appears to be a reasonable practical alternative to traditional student-based subject pools. It has several practical advantages, the main one being the available infrastructure that makes it relatively easy and painless to collect large amounts of data quickly. Mechanical Turk participants can also be truly randomly assigned to treatments, which is a methodological advantage, because it is highly impractical to do the true random assignment in the laboratory. Another advantage is that Mechanical Turk workers can be prescreened to be "qualified" for a particular study. So it is easy to conduct an experiment on only one woman or man or on participants who can correctly

answer some set of questions. Of course a potential downside is that the experimenter has no way to ensure that the participants indeed answer the questions themselves.

This brings us to some disadvantages of Mechanical Turk. While in the laboratory we can ensure that each subject participates one time only and that participants do not communicate during the study, on Mechanical Turk there is some risk of an individual creating multiple IDs or several individuals participating together. Paolacci, Chandler, and Ipeirotis (2010) consider these risks with Mechanical Turk to be relatively low, however. A bigger problem is potential self-selection bias that may happen when participants fail to complete the entire task. If the reason participants drop out has to do with the nature of the task itself, and the dropout percentage is non-negligible, or even worse, different across treatments, then this may seriously bias and confound the results. So researchers should carefully think about their study and all the trade-offs involved in different subject pools. Ultimately, experimental results should be replicated with different subject pools to understand the level of generality of the findings.

1.5 Incentives

In this section I will discuss the role of incentives. Economists use monetary incentives in their experiments. Smith (1976) introduced the idea of *induced-value theory* that explains that using monetary incentives provides a way to gain control over economically relevant characteristics of the laboratory participants. In other words, paying subjects based on their performance in the game causes them to wish to perform better because better performance results in making more money. If the amounts of money subjects earn are significant to them, and if they were recruited using earning money as the incentive (as opposed, for example, to giving course credit for participating), then the participants' innate characteristics become less relevant, and researchers can be more confident that their participants are truly trying to play the game in a way that was meant.

Financial incentives are most convenient, but in principle, other types of incentives can be used. The main factor is that the amount of the reward medium earned should be proportional to how well participants perform (as opposed to being given simply for participating). So, for example, in theory course credit could be used, as long as the amount of course credit is proportional to the amount of profit made in the game. In practice it is difficult to make course credit given in this way sufficiently salient, though.

There are a number of valid variations in incentive-compatible ways to reward participants. The *binary lottery* procedure involves awarding participants virtual lottery tickets based on their performance – each lottery ticket increases the probability of winning a prize.

This procedure has a theoretical advantage of controlling for risk aversion (because regardless of risk preferences, everyone should prefer more lottery tickets to fewer (see Roth 1995a)), but a practical disadvantage of being less straightforward than simply paying money.

Another variation is to pay for one or several randomly chosen rounds instead of the average for all rounds. Neither method can be said to be clearly better, so it is a matter of preference which payment method is used.

A more important practical question is to what extent using real incentives matters. Much of important and influential work has been based on experiments based on hypothetical choices (Kahneman and Tversky 1979), and experiments that use hypothetical choices are accepted in many branches of social science, such as psychology, marketing, and organizational behavior. Sometimes behavior in hypothetical situations does not differ from behavior in real situations, but sometimes it does differ. I will discuss two studies that directly address this issue.

Forsythe, Horowitz, and Sefton (1994) investigate the reasons for more equitable distribution in the ultimatum game (Güth, Schmittberger, and Schwarz 1982) than the subgame perfect equilibrium prediction. The authors consider two alternative hypotheses for equitable distributions: (i) Proposers are trying to be fair to responders, or (ii) proposers make large offers because they realize that responders are likely to reject offers that are too small. In order to be able to distinguish between the two hypotheses, the authors conducted some treatments with a modification of the ultimatum game, called the dictator game, the only difference being that in the dictator game responders cannot reject offers – they have to simply accept whatever (if any) offer the proposer chooses. If equitable distribution is driven primarily by the proposers' desire to treat responders fairly, the offers in the ultimatum and the dictator games should not differ. But if it is the fear of being rejected that drives equitable offers, then offers in the dictator game should be significantly lower.

The authors conducted their two games (ultimatum and dictator) under two different payment conditions: real and hypothetical. Figure 1.4 displays histograms of offers in the four treatments on the Forsythe et al. (1994) study. Each treatment included two separate sessions (April and September), and within each treatment the distributions for April and September do not differ.

The striking point is that the distributions of offers without pay are not different for the ultimatum and the dictator games (compare Figure 1.4c and d), while with pay they are strikingly different (compare Figure 1.4a and b). In other words, proposers are quite generous with hypothetical money, but not with real money. Had this study been conducted without real incentives, the researchers would have drawn incorrect conclusions about the underlying causes for equitable distributions in the ultimatum game.

Another well-known study that directly compares real and hypothetical choices is by Holt and Laury (2002). The authors study the effect of the

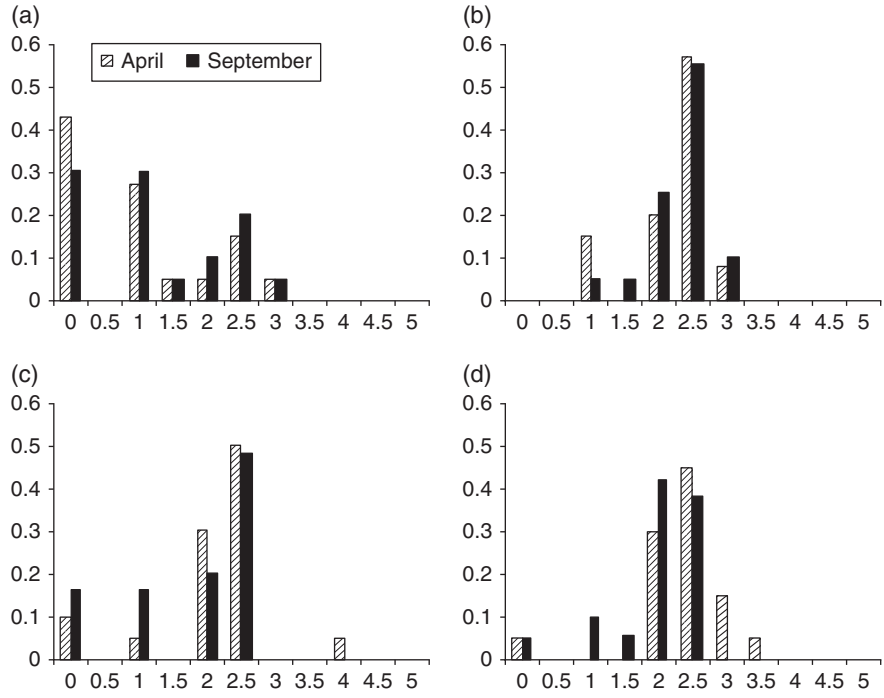

Figure 1.4 Distribution of offers in the Forsythe, Horowitz, and Sefton (1994) study. (a) Dictator game with pay, (b) ultimatum game with pay, (c) dictator game without pay, and (d) ultimatum game without pay.

magnitude and real vs. hypothetical incentives on risk preferences. The instrument they use to elicit risk preferences is presented in Table 1.1.

Participants are asked to make a choice between the Option A and Option B lottery in each row. The Option A lottery is safe, while the Option B lottery is risky. But as we move down the rows, the probability of a high payoff in the Option B lottery increases (and becomes certain in the 10th row). A risk-neutral subject should switch from Option A in row 4 to Option B in row 5, but the more risk-averse participants may switch later. Eventually every participant should prefer Option B in the 10th row.

Holt and Laury (2002) vary the magnitude of the stakes by conducting treatments with payoffs in Table 1.1 multiplied by the factors of 20, 50, and 90. They also conduct each treatment with real as well as hypothetical stakes.

Figure 1.5 shows the summary of the proportion of participants choosing Option A in each treatment. More risk-averse individuals should choose more Option A's. The key finding is that behavior looks very similar for small stakes real choices and for hypothetical choices and the size of the stakes does not

Table 1.1 The instrument to elicit risk preferences in Holt and Laury (2002).

Option A	Option B	Expected payoff difference
1/10 of \$2, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10	\$1.17
2/10 of \$2, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10	\$0.83
3/10 of \$2, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10	\$0.50
4/10 of \$2, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10	\$0.16
5/10 of \$2, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10	–\$0.18
6/10 of \$2, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10	–\$0.51
7/10 of \$2, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10	–\$0.85
8/10 of \$2, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10	–\$1.18
9/10 of \$2, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10	–\$1.52
10/10 of \$2, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10	–\$1.85

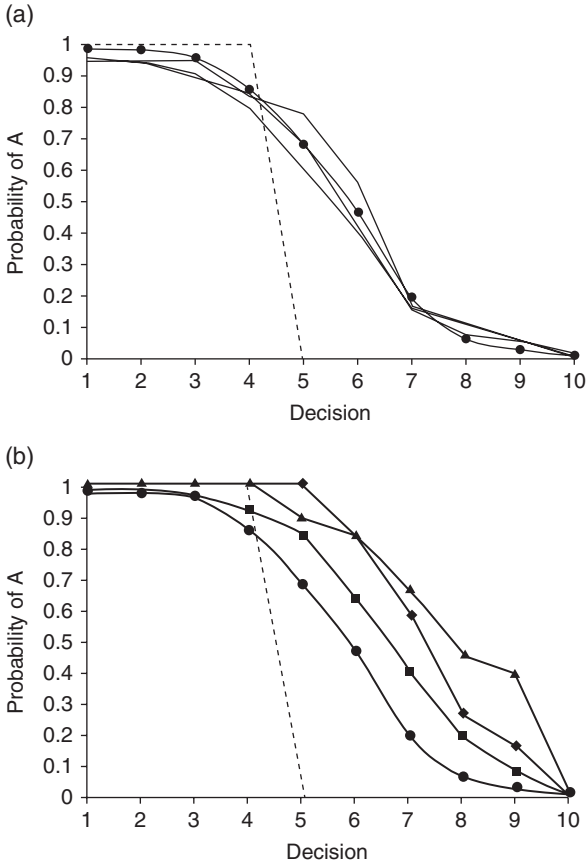

Figure 1.5 Summary of Holt and Laury (2002) data. (a) Low real payoffs (solid line with dots) compared with hypothetical payoffs (thin lines) and risk-neutral benchmark (dashed line). (b) Real payoffs: low (solid line with dots), 20× (squares), 50× (diamonds), 90× (triangles), risk neutral (dashed).

seem to matter much with hypothetical choices (Figure 1.5a). However, risk aversion levels increase as real stakes increase (Figure 1.5b).

There are other types of decisions, however, for which there is no evidence that real vs. hypothetical payments make a difference. For example, in the newsvendor experiments, the Schweitzer and Cachon (2000) study had virtually no real incentives, but when Bolton and Katok (2008) replicated the study with carefully controlled incentives, they found no discernable difference in behavior.

Another issue that should be seriously considered is that without real incentives participants may pay less attention and the resulting behavior may well be noisier. So if a study is being conducted without real incentives, it is particular important that there is some other mechanism to ensure that participants take their decisions seriously. For example, participants might be asked for their help with research and be given an option to opt out. Those are, admittedly, weak manipulations – providing real incentives is, as a rule, better.

In summary, we can conclude that providing real incentives can matter a great deal. Decisions that involve social preferences or risk are definitely affected by real incentives. Decisions that involve more straightforward optimization tasks seem to be less affected by real incentives. It is not always clear a priori whether real incentives will matter or not. Therefore, initial experiments should be conducted with real incentives, until it has been systematically shown that behavior is not affected by hypothetical incentives.

1.6 Deception

The next methodological topic I will discuss is the use of deception. What constitutes deception ranges from deliberately providing subjects with false information to not specifically stating some information, allowing subjects to draw their own, perhaps incorrect, conclusions. An example of the former is telling participants that they are interacting with another human participant, while in fact they are interacting with a computer. An example of the latter might be inviting 32 participants into the lab, matching them repeatedly in four group of eight, but only telling them that each period they are randomly matched with another person in the room (a technically true but slightly misleading statement). While both are examples of deception, the former is definitely considered unacceptable by experimental economists, while the latter is not.

Davis and Holt (1993) cite the loss of experimental control, as the primary reason deception is considered unacceptable. Most economists are concerned about developing and maintaining a reputation among the student population for honesty in order to ensure that subject actions are motivated by the induced monetary rewards rather than by psychological reactions to suspected manipulation (pp. 23–24). This point again comes down to maintaining experimental control.

There are two ways experimental control might suffer due to the use of deception: indirect and direct. Participants who have experienced deception in previous experiments may not trust the experimenters in future, unrelated experiments. Thus, the use of deception by a few researchers might (indirectly) contaminate the entire subject pool. There is also potentially a direct loss of control, because, when subjects are being deceived in a study, they may (correctly) suspect that they are being deceived. After all, a reason to deceive subjects in the first place is to investigate phenomena that may not naturally occur without deception. It is difficult to assess the direct effect of deception, but generally speaking, since deception diminishes control, it is better to try to design experiments without using deception.

The use of deception is common in psychology. In their review article, Ortmann and Hertwig (2002) report that more than 1/3 of studies published in psychology use deception. Even more importantly, studies that use deception are routinely studied in undergraduate psychology courses. Since the subject pool for psychology studies typically comes from the population of undergraduate psychology majors, these participants are generally aware that they are likely to be deceived, and they tend to expect this. Moreover, the type of deception psychology studies use often includes directly deceiving subjects about the purpose of the experiment, or using confederates, and investigating resulting behavior. Jamison, Karlan, and Schechter (2008) provide the following typical example of deception in a psychology study: "…subjects were given two poems by Robert Frost, were told that one was by Frost and one by a high school English teacher, and were then asked to rate the merits of the two poems. After the experiment they were debriefed and told that both poems were actually by Frost and that the experiment was looking at how beliefs regarding authorship affected the rating of the poems" (p. 478).

In contrast, experimental economists almost never use deception, so participants in economic experiments do not generally expect to be deceived. In a few studies that used deception, they tend to use it for convenience or to study reactions to behavior that is unlikely to occur naturally. This effect is usually achieved by telling subjects that they are matched with a human participant, while they are in fact matched with a computer agent programmed to behave in some specific way (Weimann 1994; Blount 1995; Scharlemann et al. 2001; Sanfey et al. 2003; Winter and Zamir 2005).

There are some studies that have investigated indirect effects of deception. Jamison, Karlan, and Schechter (2008) are perhaps the most direct study. The authors conducted an experiment that consisted of two parts. During the first part, participants played the trust game.[3] Half of the participants were not

3 In the trust game the first mover must decide on the amount x of her initial endowment to pass to player 2. This amount triples, and player 2 decides on the amount $y \leq 3x$ to return to player 1.

deceived, and the other half were deceived in that they were told that they were matched with a human participant, while in fact they were matched with a computer programmed to imitate the behavior of human participants in earlier studies. The deceived participants were debriefed at the end of the study and told that they were matched with computerized partners.

Three weeks later the authors conducted the second phase of the study, for which they invited the same group of participants to an experiment that looked unrelated. This second experiment involved a dictator game, a risk version assessment task similar to Holt and Laury (2002), and a prisoner dilemma game. Jamison, Karlan, and Schechter (2008) analyzed the effect of having been previously deceived on participation rates in the second study and on the behavior in the second study.

Jamison, Karlan, and Schechter (2008) report that deception does have an effect on participation as well as behavior. Females who have been deceived are significantly less likely to return than the females who have not been. Also, participants who were unlucky and have been deceived are less likely to return than the participants who have been unlucky but have not been deceived. In terms of behavior, participants who have been deceived behave more erratically (less consistently) in answering the risk aversion questions, indicating that they may not be taking the study as seriously. The only other difference between deceived and not deceived participants is that females or inexperienced subjects who have been deceived, and who had the role of the first moved in the trust game, tend to give less in the dictator game.

One may argue that the evidence we have so far indicates that the indirect effects of deception in terms of damaging the subject pool seem to be fairly minor. It may be that the true costs are actually higher, because the participants in the Jamison, Karlan, and Schechter (2008) study came from the economics subject pool, so they were students who were not previously deceived. A single deception incident may not have significantly changed their behavior, but it may be that repeatedly deceiving participants will alter the characteristics of the subject pool in more serious and permanent ways (see Roth 2001 for a related argument).

1.7 Collecting Additional Information

It is usually a good idea to collect demographic information from participants. Typically this information includes gender, major, degree status (undergraduate, MS, PhD), possibly year of expected graduation, and age. This information is good to have in case questions about composition of the subject pool arise during the review process, or, possibly, new research questions arise in regard to correlations between demographics and observed behavior.

Another type of information researchers sometimes collect has to do with obtaining independent measures of participants' utility functions. The most common type of measure of this kind is risk preferences. The idea is to estimate risk preferences using an independent instrument and then check to what extent these risk preferences correlate with behavior observed in the experiment.

Unfortunately, even when risk aversion could potentially explain observed behavior, risk preferences measured using a separate instrument do not correlate with the behavior. For example, Isaac and James (2000) estimated risk preferences using the Becker–DeGroot–Marschak (BDM) procedure, as well as the sealed-bid first price auction, and found negative correlation between the two measures. This general failure to find positive correlation across risk preferences measured using different tasks may be because risk preferences themselves are not stable across tasks, or because behavior is really not explained by risk preferences, or because the instrument used to measure risk preferences is not reliable. In fact, how to reliably measure risk preferences remains an open question.

Crosetto and Filippin (2016) compares four common incentivized risk elicitation tasks, as well as two nonincentivized questionnaire-type tasks, and find that the measures the four tasks yield are not consistent. The incentivized tasks are Holt and Laury (2002) (already discussed in the previous section) and the Eckel and Grossman (2002) task that consists of choosing one lottery from a list of five ordered in terms of increased expected value and variance. The main difference between the Holt and Laury (2002) and the Eckel and Grossman (2002) is that the Holt and Laury instrument keeps payoffs of the two lotteries the same and changes probabilities, while the Eckel and Grossman (2002) instrument keeps probabilities at 50/50 and manipulates the payoffs. Also, the Eckel and Grossman (2002) instrument requires a single decision (select the preferred lottery), while the Holt and Laury (2002) requires a decision for each lottery pair, introducing potential for errors. The third task is the investment game (Gneezy and Potters (1997) in which a participant is asked to allocate four euro between a safe account and a lottery with a 50/50 chance of earning either zero of 2.5 times the amount allocated to the lottery. The fourth is the Bomb Risk Elicitation Task (BRET) by Crosetto and Filippin (2016), in which subjects are faced with 100 "boxes" of which one contains a "bomb" and are asked how many of the boxes to open. If one of the open boxes contains the "bomb," the earnings are zero, and otherwise the earnings are proportional to the number of opened boxes. The two additional questionnaire-type instruments are the German Socio-Economic Panel Study (SOEP) risk question (Wagner, Frick, and Schupp 2007) and the Domain-Specific Risk-Taking Scale (DOSPERT) (Blais and Weber 2006). The main finding Crosetto and Filippin (2016) report is that the Eckel and Grossman (2002) task produced a significantly higher measure of risk aversion than the other tasks did, followed by the

Holt and Laury (2002). Additionally, about 20% of subjects made inconsistent choices in the Holt and Laury (2002) task (switched more than once). Interestingly, gender differences were significant in the Eckel and Grossman (2002) and the investment tasks, but not in the other two. The bottom line is that the elicitation of risk preferences is quite unreliable, and not finding correlation between elicited risk preferences and the observed behavior is not unusual.

Some other types of measures that researchers sometimes collect include CRT (Frederick 2005), which is designed to measure the ability to suppress an intuitive wrong answer to come up with a more deliberate correct answer, or measuring numeracy (Cokely et al. 2012). Moritz, Hill, and Donohue (2013) report that the CRT score is correlated with the performance in the newsvendor game.

1.8 Infrastructure and Logistics

Some of the infrastructure and logistic requirements needed to conduct laboratory experiments include funding to pay the participants, an efficient way to recruit those participants, the approval for the use of human subjects that is required by US universities, the software to implement the games, and a computer lab in which to conduct the study.

Laboratory experiments tend to be relatively inexpensive compared, for example, with experiments conducted in natural of physical sciences. Many research-oriented universities provide small grants for data collection that is often sufficient for a study with a reasonable sample size.

Subject recruitment is most efficiently done through the Internet, and several recruitment systems have been developed and are freely available for academic use (ORSEE software, developed by Ben Greiner, can be accessed from this URL: http://www.orsee.org/). Psychology departments often use Sona Systems (http://www.sona-systems.com), which is not free, but is convenient, because it is hosted.

Human subject approval typically requires providing information about your study and your recruitment process to an office on campus (this is usually the same office that reviews medical studies to ensure that human subjects are not subjected to major risks). Studies involving laboratory experiments in social studies also have to be reviewed by the Institutional Review Board (IRB) or a similar regulatory body. Check your university rules about obtaining human subject approval for studies before you start your work.

The majority, although not all, of experiments are conducted using a computer interface. Computer interface is a convenient and efficient way to collect data, but the downside is that implementing even a simple game may require a significant amount of work. Fortunately, there are several systems that have

been designed to simplify implementing simple games. These systems have a fairly intuitive structure and syntax that is easy to learn even for a person with modest programming skills. The oldest platform was developed by Urs Fischbacher and is called z-Tree (Zurich Toolbox for Ready-made Economic Experiments) for implementing laboratory experiments (http://www.iew.uzh.ch/ztree/index.php). This software is freely available to academic researchers. And it has a good tutorial, a wiki, and an active user listserv. z-Tree is designed to be used in a computer lab on computers networked using a LAN. It is flexible enough to create experimental software quickly and even includes some advanced GUI features, such as graphs and chat boxes (see Fischbacher 2007). The main downside of z-Tree is that it cannot be used over the Internet.

Several new web-based platforms have been recently developed, including oTree (http://www.otree.org) and SoPHIE (Software Platform for Human Interaction Experiments; http://www.sophielabs.com). SoPHIE also includes a convenient GUI for implementing simple experiments, as well as additional commercial tools for more complex, real-time experiments (bargaining, auctions, markets, chatting) (see Hendriks 2012).

z-Tree can be easily installed in any computer lab, and SoPHIE and oTree only require a web browser, so a dedicated lab, although convenient and useful to have, is not essential. If you are fortunate enough to be given access to a dedicated lab, some useful features to have are privacy partitions for subject computers and an overhead projector. Larger labs are more convenient because they facilitate larger sessions, making data collection more efficient.

Using computer interface to conduct laboratory experiments is a common practice, and the ease and availability of the z-Tree and SoPHIE and oTree software platforms contributed to this trend. Computer interface usually increases control and greatly simplifies the logistics of collecting data. Occasionally, though, designs require experiments to be conducted by hand. One example of this need to run an experiment without a computer is a paper by Bolton and Zwick (1995) that studies the effect of participant–experimenter anonymity. The authors needed a transparent way to ensure not only true anonymity but also the perception of anonymity in a type of a simplified ultimatum game. They achieved this effect by conducting the experiment by hand, which involved participants passing marked boxes back and forth with the aid of several research assistants.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risqe: critique des postulats et axioms de l'ecole americane. *Econometrica* 21: 503–546.

Atkinson, A.C. and Donev, A.N. (1992). *Optimum Experimental Designs*. Oxford, England: Clarendon Press.

Alexrod, R. (1984). *The Evolution of Cooperation*. Basic Books.
Ball, S.B. and Cech, P. (1996). Subject pool choice and treatment effects in economic laboratory research. *Experimental Economics* 6: 239–292.
Blais, A.-R. and Weber, E.U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making* 1: 33–47.
Blount, S. (1995). When social outcomes aren't fair: the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63: 131–144.
Bolton, G.E. (1991). A comparative model of bargaining: theory and evidence. *American Economic Review, American Economic Association* 81 (5): 1096–1136.
Bolton, G.E. and Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior* 10 (1): 95–121.
Bolton, G. and Katok, E. (2008). Learning-by-doing in the newsvendor problem: a laboratory investigation. *Manufacturing and Service Operations Management* 10 (3): 519–538.
Bolton, G. and Ockenfels, A. (2000). A theory of equity, reciprocity, and competition. *American Economics Review* 90 (1): 166–193.
Bolton, G.E., Ockenfels, A., and Thonemann, U. (2012). Managers and students as newsvendors. *Management Science* 58 (12): 2225–2233.
Camerer, C. (1995). Individual decision making. In: *The Handbook of Experimental Economics*, vol. 1 (ed. J.H. Kagel and A.E. Roth), 587–704. Princeton University Press.
Camerer, C.F., Dreber, A., Forsell, E. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* doi: 10.1126/science.aaf0918.
Chamberlin, E.H. (1948). An experimental imperfect market. *Journal of Political Economy* 56 (2): 95–108.
Cokely, E.T., Galesic, M., Schulz, E., and Ghazal, S. (2012). Measuring risk literacy: the Berlin numeracy test. *Judgment and Decision Making* 7 (1): 25–47.
Cooper, D.J. and Kagel, J.H. (2016). Other-regarding preferences: a selective survey of experimental results, 1995–2008. In: *The Handbook of Experimental Economics*, vol. 2 (ed. J.H. Kagel and A.E. Roth). Princeton University Press.
Crosetto, P. and Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics* 19 (3): 613–541.
Cui, T.H., Raju, J.S., and Zhang, Z.J. (2007). Fairness and channel coordination. *Management Science* 53 (8): 1303–1314.
Davis, D.D. and Holt, C.A. (1993). *Experimental Economics*. Princeton: Princeton University Press.
Eckel, C.C. and Grossman, P.J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior* 23 (4): 281–295.
Fehr, E. and Schmidt, K.M. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114 (3): 817–868.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10 (2): 171–178.
Flood, M.M. (1958). Some experimental games. *Management Science* 5: 5–26.
Fisher, R.A. (1935). *The Design of Experiments Games*. Edinburgh, Scotland: Oliver and Boyd.
Forsythe, R., Horowitz, S., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior* 6 (3): 347–369.
Fréchette, G.R. (2012). Laboratory experiments: professionals versus students. In: *The Methods of Modern Experimental Economics* (ed. G. Frechette and A. Schotter). Oxford: Oxford University Press.
Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19 (4): 25–42.
Friedman, D. and Sunder, S. (1994). *Experimental Methods a Primer for Economists*. Cambridge University Press.
Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics* 112 (2): 631–645.
Goeree, J.K. and Halt, C.A. (2009). Hierarchical package bidding: a paper & pencil combinatorial auction. *Games and Economic Behavior* 70 (1): 146–169.
Güth, W., Schmittberger, R., and Schwarz, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3 (4): 367–388.
Hendriks, A. (2012). SoPHIE – software platform for human interaction experiments. Working paper. University of Osnabrueck.
Holt, C.A. (1995). Industrial organization: a survey of laboratory results. In: *Handbook of Experimental Economics* (ed. J. Kagel and A. Roth), 349–443. Princeton: Princeton University Press.
Holt, C.A. and Laury, S.K. (2002). Risk aversion and incentive effects. *American Economic Review* 92 (5): 1644–1655.
Isaac, R.M. and James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty* 20: 177–187.
Jamison, J., Karlan, D., and Schechter, L. (2008). To deceive or not to deceive: the effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization* 68: 477–488.
Kagel, J.H. (1995). Auctions: a survey of experimental research. In: *The Handbook of Experimental Economics* (ed. J.H. Kagel and A.E. Roth), 501–585. Princeton: Princeton University Press.
Kagel, J. and Levin, D. (1986). The winners curse and public information in common value auctions. *American Economic Review* 76 (5): 894–920.
Kagel, J.H. and Levin, D. (2016). Auctions: a survey of experimental research, 1995–2008. In: *The Handbook of Experimental Economics*, vol. 2 (ed. J.H. Kagel and A.E. Roth). Princeton University Press.
Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47: 263–291.

Katok, E. and Siemsen, E. (2011). The influence of career concerns on task choice: experimental evidence. *Management Science* 57 (6): 1042–1054.

Katok, E., Thomas, D., and Davis, A. (2008). Inventory service level agreements as coordination mechanisms: the effect of review periods. *Manufacturing & Service Operations Management* 10 (4): 609–624.

Krishna, V. (2002). *Auction Theory*, 1e. San Diego: Academic Press.

Lee, Y.S., Seo, Y.W., and Siemsen, E. (2017). Running behavioral operations experiments using Amazon's Mechanical Turk. *Production and Operations Management* doi: 10.1111/poms.12841.

List, J.A., Sadoff, S., and Wagner, M. (2011). *Experimental Economics* 14: 439. https://doi.org/10.1007/s10683-011-9275-7.

Loch, C.H. and Wu, Y. (2008). Social preferences and supply chain performance: an experimental study. *Management Science* 54 (11): 1835–1849.

Machina, M. (1997). Choice under uncertainty: problems solved and unsolved. *Journal of Economic Perspectives* 1 (1): 121–154.

Moritz, B.B., Hill, A.V., and Donohue, K. (2013). Cognition and individual difference in the newsvendor problem: behavior under dual process theory. *Journal of Operations Management* 31 (1-2): 72–85.

Ochs, J. and Roth, A.E. (1989). An experimental study of sequential bargaining. *American Economic Review* 79: 355–384.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349 (6251).

Ortmann, A. and Hertwig, R. (2002). The costs of deception: evidence from psychology. *Experimental Economics* 5: 111–131.

Özer, Ö., Zheng, Y., and Chen, K. (2011). Trust in forecast information sharing. *Management Science.* 57 (6): 1111–1137.

Paolacci, G., Chandler, J., and Ipeirotis, P.G. (2010). Running experiments on Amazon mechanical Turk. *Judgment and Decision Making* 5 (5): 411–419.

Plott, C. (1987). Dimensions of parallelism: some policy applications of experimental methods. In: *Experimental Economics: Six Points of View* (ed. A. Roth). New York: Cambridge University Press.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: an experimental study. *The American Economic Review* 81 (5): 1068–1095.

Roth, A.E. (1984). The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy* 92 (1984): 991–1016.

Roth, A.E. (1995a). Introduction to experimental economics. In: *The Handbook of Experimental Economics*, vol. 1 (ed. J.H. Kagel and A.E. Roth), 3–109. Princeton University Press.

Roth, A.E. (1995b). Bargaining experiments. In: *The Handbook of Experimental Economics*, vol. 1 (ed. J.H. Kagel and A.E. Roth), 253–248. Princeton University Press.

Roth, A.E. (2001). Form and function in experimental design. *Behavioral and Brain Sciences* 24: 427–428.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica* 50 (1): 97–109.

Sanfey, A.G., Rilling, J.K., Aronson, J.A. et al. (2003). The neural basis of economic decision-making in the ultimatum game. *Science* 300: 1755–1758.

Scharlemann, J.P.W., Eckel, C.C., Kacelnik, A., and Wilson, R.K. (2001). The value of a smile: game theory with a human face. *Journal of Economic Psychology* 22: 617–640.

Schweitzer, M. and Cachon, G. (2000). Decision bias in the newsvendor problem: experimental evidence. *Management Science* 46 (3): 404–420.

Smith, V.L. (1962). An experimental study of competitive market behavior. *The Journal of Political Economy* 70 (2): 111–137.

Smith, V.L. (1976). Experimental economics: induced value theory. *American Economic Review* 66 (2): 274–279.

Smith, V.L. (1982). Microeconomic systems as an experimental science. *American Economic Review* 72: 923–955.

Sterman, J. (1989). Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* 35 (3): 321–339.

Thurstone, L.L. (1931). The indifference function. *Journal of Social Psychology* 2: 139–167.

Tversky, A. and Kahneman, D. (1981). The framing of decisions in the psychology of choice. *Science* 211 (4481): 453–458.

Von Neumann, L. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wagner, G.G., Frick, J.R. and Schupp, J. (2007). The German socio-economic panel study (SOEP): scope, evolution and enhancements. SOEP papers on Multidisciplinary Panel Data Research 1, DIW Berlin, The German Socio-Economic Panel (SOEP).

Wallis, W.A. and Friedman, M. (1942). The empirical derivation of indifference functions. In: *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz* (ed. O. Lange, F. McIntyre and T.O. Yntema), 175–189. Chicago, IL: Chicago University Press.

Weimann, J. (1994). Individual behavior in a free riding experiment. *Journal of Public Economics* 54: 185–200.

Winter, E. and Zamir, S. (2005). An experiment on the ultimatum bargaining in a changing environment. *Japanese Economic Review* 56: 363–385.