INFORMS—Charlotte 2011

 | ISBN 0000-0000
DOI 10.1287/educ.1053.0000

Chapter 2

Laboratory Experiments in Operations Management

Elena Katok
Penn State University, Smeal College of Business,
University Park, PA 16870, USA
ekatok@psu.edu

Abstract Controlled laboratory experiments give researchers a great deal of control in making them useful for testing analytical models. In this tutorial I introduce laboratory experiments and discuss methodological issues in designing and conducting laboratory experiments.

Keywords Operations management; behavioral economics

1. Introduction

Much of the work in Behavioral Operations Management (BOM) lives at the boundary of analytical and behavioral disciplines—work that has a substantial tradition. In the next section I will briefly summarize the history of the uses of laboratory experiments in economics, and how the field of BOM can learn from this tradition.

Laboratory experiments are a major method we use in BOM. Similar methods have been employed in a number of other social science fields, including for example economics (auctions), psychology and sociology (social networks), law (jury behavior), political science (coalition formation), anthropology and biology (reciprocity).

There are three major purposes that laboratory experiments serve (Roth 1995a). (1) To test and refine existing theory. Much of the BOM work so far has been on this topic. For example, experiments testing behavior in the newsvendor model (Schweitzer and Cachon 2000, Bolton and Katok 2008). (2) To characterize new phenomena leading to new theory. An excellent example is the literature on social preferences (Loch and Wu 2008, Cui et al. 2007). (3) To test new institutional designs. This type of work has not yet made its way in the operations literature, but there are several notable examples in economics, such as designing the FCC auctions for radio spectrum (Goeree and Holt 2010) or designing the market for medical interns (Roth 1984).

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. 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. 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.

The relative advantage of laboratory experiments is control. Experiments can be designed to fully manipulate all factors at all desired levels, and to match the assumptions of the analytical model being tested. So laboratory experiments are high on the internal validity, but because the environment is often more artificial, they are lower on the external validity.

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 fractile 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—a 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 fractile 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, bargainers may assume to 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 (Rubenstein 1982) that provides useful insights and 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 (see Ochs and Roth 1989, Bolton 1991) and the results of these tests provided seminal insights that formed the basis for the theory of social preferences (Bolton and Ockenfels 2000, Fehr and Schmidt 1999).

One of the questions often asked about laboratory experiments is about 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 (1) it

has empirical implications, and (2) 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.

Some of the other elements 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 game, and a computer lab to run the experiment.

Laboratory experiments tend to be relatively inexpensive compared, for example, to experiments conducted in natural or 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/).

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 insure that human subjects are not subjected to major risks). Studies involving laboratory experiments in social studies unfortunately also have to be reviewed. Check your university rules about obtaining human subject approval for studies before you start your work.

The majority, although not all, experiments are conducted using a computer interface. A 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, Urs Fischbacher developed a platform called z-Tree (Zurich Toolbox for Readymade Economic Experiments) for implementing laboratory experiments (http://www.iew.uzh.ch/ztree/index.php). This software is freely available to academic researchers. It has a fairly intuitive structure and syntax that is easy to learn even for a person with modest programming skills. 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, so it is not ideal for the use over the internet, but this is perhaps its only limitation. 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).

z-Tree can be easily installed in any computer lab, 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.

The rest of this article is organized as follows: in Section 2, I will present a (very) short history of experimental economics, focusing specifically on some fundamental games that proved to be important in economics as well as in BOM. These games will come up again in subsequent sections. In Section 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 article in Section 4 with a discussion of my view of future trends and promising directions for future research.

2. A Historical Perspective

2.1. 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. 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 a heads comes up. 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, \$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 $\sum_{n=1}^{\infty} \frac{n}{2^n}$, 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 individuals' 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 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 alternative 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 D to C, 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 (1987) for an overview and Camerer (1995) for a detailed literature survey of individual decision-making.

In spite of numerous documented violations, expected utility theory continues to be the predominant paradigm in economics. 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 Operations Management, in spite of Bernoulli's early demonstration in 1728, the majority of models assume risk neutrality, and even allowing for risk aversion is a fairly new phenomenon.

2.2. Simple Strategic Games

Following von Neumann and Morgenstern (1944), economists also became interested in testing models of strategic interactions. 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

FIGURE 1. Payoffs in the Prisoner's Dilemma Game (Flood 1958).

		Column Player: Defect	Column Player: Cooperate
Row Player	Cooperate	Row Earns -1 Column Earns 2	Row Earns ½ Column Earns 1
	Defect	Row Earns 0 Column Earns ½	Row Earns 1 Column Earns -1

label Cooperate and Defect, but that carried neutral labels "1" and "2" in the experiments). The payoffs are displayed in Figure 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 1/2 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. 1/2 if Column Cooperates, and 0 vs. -1 if Column Defects). Thus the unique equilibrium in the Prisoner's Dilemma game is for both players to defect, Row earning 0 and Column earning 1/2. 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 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 Operations Management is bargaining. Güth, Schmittberger and Schwarze (1982) were the first to conduct an experiment on the *Ultimatum Game*, that has since become the standard vehicle for modeling the negotiation process. The game involves two players. The Proposer receives $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 a unique subgame perfect equilibrium that can be computed using *backwards 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 et al. (1982) experiments were conducted, hundreds of ultimatum experiments have been reported. Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991) conducted a large-scale study in four countries: US, 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 equilibrium prediction, in which 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 (Axelrod 1984).

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 backwards 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 important 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, and (2) the deadline effect: most agreements happen in the last second. In regards 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, Sefton and Savin (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 some ground. Contributions reflect "pure" preferences. I will discuss the Forsythe et al. (1994) paper in more detail in the following section. I refer the reader to Roth (1995b) for a review of bargaining experiments prior to 1995. These early laboratory experiments also gave rise to both analytical and behavioral literature on other-regarding preferences. I refer the reader to Cooper and Kagel (2008) for a review.

2.3. 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).

[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).

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), and laboratory experiments have been used to test many of these models. I refer the reader to Kagel (1995) for a comprehensive review of work done prior to 1995 and to and Kagel and Levin (2008) for work done since 1995.

There are two streams of research that are particularly relevant to BOM that I will briefly summarize here. The first deals with testing the model in which bidders are assumed to have valuations that are independent, drawn from the same distribution (symmetric), and privately known (the independently-known private value (IPV) model).

Much of the early laboratory experiments on auctions dealt with testing the implications of the revenue equivalence theory (for forward auctions) in the IPV setting. Vickrey (1961) showed that if bidders are risk neutral, the expected seller revenues in forward auctions are the same in the four basic auction formats:

(1) *The sealed-bid first price*: bidders submit sealed bids and the object is awarded to the bidder who submitted the best bid, and this bidder pays his bid.
(2) *The sealed-bid second price*: bidders submit sealed bids and the object is awarded to the bidder who submitted the best bid, but he pays the amount of the second best bid.
(3) *The open-bid ascending* (*English*): bidders place bids dynamically during a live event. At the end of the auction the object is awarded to the bidder who submitted the best bid, and he pays the amount of his bid.
(4) *Clock descending* (*Dutch*): the price starts high and decreases at a regular pre-determined rate (the clock). The first bidder to stop the clock wins the object and pays the price on the clock.

If bidders are not risk neutral, the equivalence does not generally hold. If they are risk averse the equivalence holds between the sealed-bid first price and Dutch, and the sealed-bid second price and English.

Virtually all laboratory work to date that deals with revenue equivalence in auctions deals with forward auctions (see Kagel 1995 for a review). Generally, laboratory tests reject all versions of revenue equivalence. Sealed-bid first price revenues were reported to be higher than Dutch revenues (Cox, Roberson & Smith 1982), but later Lucking-Reiley (1999) reported the opposite effect in a field experiment. Katok and Kwasnica (2008) found that prices in the Dutch auction critically depend on the speed of the clock, and thus can be either above or below the sealed-bid first price prices. Several models of bidder impatience have been offered (Carare and Rothkopf 2005 and Katok and Kwasnica 2008).

Similarly, there is no support for the revenue equivalence between the sealed-bid second price and English auctions because, although both formats have the same dominant bidding strategy, bidders in English auctions tend to follow it, while bidders in sealed-bid second price auctions tend to place bids above their valuations (Kagel, Harstad and Levin 1987) and are extremely slow to learn to not do that.

There is an important literature stream that examines bidding behavior in sealed-bid first price auctions and compares it to the equilibrium bidding behavior. Cox, Smith and Walker (1988) reported that bidding in sealed-bid first price auctions is more aggressive than it should be in equilibrium, and thus the revenue is higher when the sealed-bid first price auction is used than when the English auction is used. Cox et al. (1988) show that qualitatively, this difference is consistent with risk aversion. Equivalently, in a procurement setting, Holt (1980) shows that when bidders are risk averse, the expected procurement cost in equilibrium is lower in sealed-bid first price auctions than in their open-descending counterparts. But as Kagel (1995) points out, in sealed-bid first-price auctions, "... risk

aversion is one element, but far from the only element, generating bidding above the ...[risk-neutral Nash equilibrium]." (p. 525).

There are a number of studies that show that risk aversion does not organize the data well in many auction-like settings (Kagel, Harstad and Levin 1987, Kagel and Levin 1993, Cason 1995, Isaac and James 2000). There is also a number of more recent studies that propose other explanations, such as aversion to regret (Engelbrecht-Wiggans and Katok 2007, Engelbrecht-Wiggans and Katok 2008, Filiz-Ozbay and Ozbay 2007), learning (Ockenfels and Selten 2005, Neugebauer and Selten 2006), and simply reacting to errors (Goeree, Holt and Palfrey 2002). While the precise explanation for overly-aggressive bidding in sealed-bid first price auctions appears to be elusive, the fact that bidders tend to bid more competitively in sealed bid than in open-bid auctions appears to be quite robust and general. In Section 6 I will show that this regularity applies to a wider set of auctions than just sealed-bid first price; the "sealed-bid effect" applies also to dynamic auctions in which bidders are not certain whether they are winning or losing the auction.

3. Established Good Practices for Conducting BOM Laboratory Experiments

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

3.1. Effective Experimental Design

In laboratory experiments, researchers generate their own data, and this allows for much 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 (1992) 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 variables*. 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 a 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, that 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*. 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 2:

FIGURE 2. An example of a 2×2 full factorial design.

		Number of Bidders	
		n=2	n=4
Auction Format	Open-Bid	OB-2	OB-4
	Sealed-Bid	SB-2	SB-4

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 effects. But the disadvantage is that in an experiment with a large number of focal variables, a full factorial design can become prohibitively expensive.

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. This yields a $2 \times 2 \times 2$ full factorial design with the following eight treatments:

$$+++ \quad ++- \quad +-+ \quad +-- \quad -++ \quad -+- \quad --+ \quad ---$$

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).

$$+++ \quad +-- \quad -+- \quad --+$$

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:

$$++++ \quad -+++ \quad +-++ \quad ++-+ \quad +++-$$

That investigated the effect of each of the four variables, and compares them to the baseline (++++) one at a time.

Some nuisance variables cannot be directly controlled (for example, subjects' 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 subjects arrive.

A *random block* design holds one or more nuisance variables constant across treatments. An example is a *within-subjects 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 subjects design introduces potential *order effects*: 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 convincing evidence that they are not there, but only that the design does not have sufficient power to detect them.

A very clever way to use within subjects 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, 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.

3.2. 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 "Mover 1" and "Mover 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) biasing 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 the same basic setting may represent just as well. So one school of thought is that since 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 a great deal of evidence that *framing* (how the problem is described to participants) can have a large effect on behavior (Kahnemann and Tversky 1979, Machina 1987). In BOM we tend to have a cover story that is related to the application we are investigating. Researchers should take great care, however, in balancing the need for context with unintentional framing and leading.

3.3. 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 there is no systematic evidence that managers perform any better (or any worse, for that matter) than do students.

There are some obvious practical reasons for using undergraduate 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 hr.). 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 my opinion, the main downside of using managers in experiments 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 weighted against the possible benefits of having a non-student subject pool.

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. 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 (2010) 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 (2010). In the context of the newsvendor game, 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 decisions, 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 3. Mean order quantities in the Bolton, Ockenfels and Thonemann (2010) experiment.

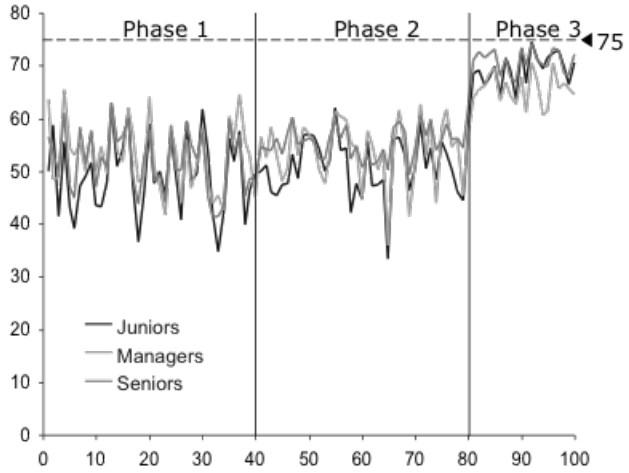


Figure 3 summarizes mean order quantities in the Bolton, Ockenfels and Thonemann (2010) study. All three groups exhibit the pull-to-center effect, and do not exhibit any learning within each phase, and all three groups performed better after the tutorial on computing the optimal order quantity (Phase 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 do seniors. 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, 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 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 a student subject pool yields different results than professionals. Therefore, using student participants in laboratory experiments is a good procedure, and is certainly a reasonable first step, unless there are some very specific reasons to believe that professionals are likely to behave significantly and systematically different.

3.4. Incentives

In this subsection I will discuss the role of incentives. Economists use real 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 Tverky 1979), and experiments that use hypothetical choices are accepted in many branches of social science, such as psychology, marketing, and management. 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, Sefton and Savin (1994) investigate the reasons for more equitable distribution in the Ultimatum game (Güth et al. 1982) than the subgame perfect equilibrium prediction. The authors consider two alternative hypotheses for equitable distributions: (1) proposers are trying to be fair to responders, or (2) 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 4 displays histograms of offers in the four treatments in 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 4 (c) and (d)), while with pay they are strikingly different (compare Figure 4 (a) 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 magnitude and real vs. hypothetical incentives on risk preferences. The instrument they use to elicit risk preferences is presented in Table 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 10^{th} 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 10^{th} row.

FIGURE 4. Distribution of offers in the Forsythe et al. (1994) study.

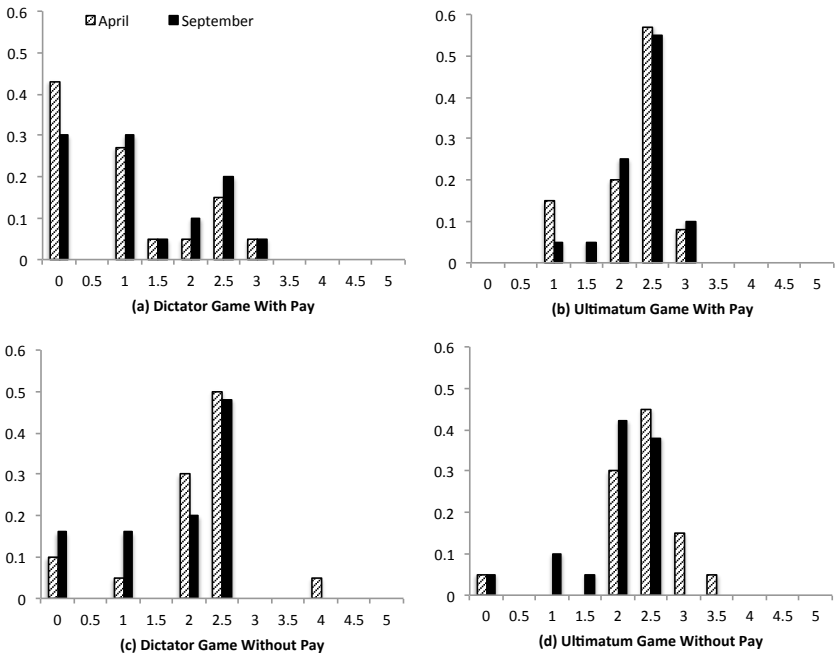


TABLE 1. The instrument to elicit risk preferences in Holt and Laury (2002).

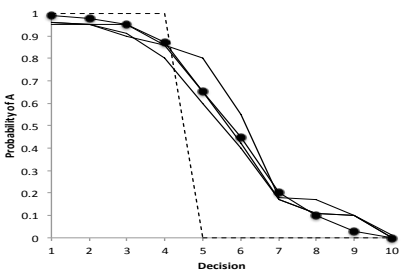
Option A	Option B	Expected payoff difference
1/10 of \$2, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10	\$1.17
2/10 of \$2, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10	\$0.83
3/10 of \$2, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10	\$0.50
4/10 of \$2, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10	\$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

Holt and Laury (2002) vary the magnitude of the stakes by conducting treatments with payoffs in Table 1 multiplied by the factors of 20, 50 and 90. They also conduct each treatment with real as well as hypothetical stakes.

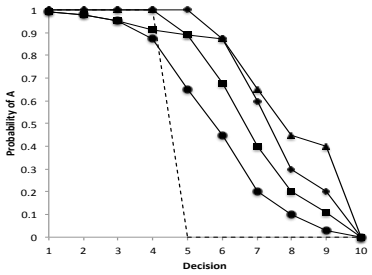
Figure 5 shows the summary of the proportion of participants choosing Option A in each treatment. More risk averse individuals should choose more Option A's. The key finding is that behavior looks very similar for small stakes and real choices and for hypothetical choices, and the size of the stakes does not seem to matter much with hypothetical choices (Figure 5(a)). However, risk aversion levels increase as real stakes increase (Figure 5(b)).

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

FIGURE 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], 20x [squares], 50x [diamonds], 90x [triangles], risk neutral [dashed]

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 particularly 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.

3.5. Deception

The last 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 groups 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 very 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).

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, Ortman 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. The few published studies that used deception, seem to have 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), and 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 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 de-briefed 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 aversion assessment task similar to Holt and Laury (2002), and a prisoner dilemma game. Jamison et al. (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 et al. (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

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

participants is that females or inexperienced subjects who have been deceived, and who had the role of the first mover 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 et al. (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).

4. Conclusion

In this section I conclude the article with some of my personal thoughts about the future directions and trends in the BOM field. Let us start by briefly looking back to the three purposes of laboratory experiments I mentioned in the introduction, and ask "how have we done so far?". The three purposes are: (1) To test and refine existing theories; (2) To characterize new phenomena leading to new theory; (3) To test new institutional designs.

Much of the effort up to this point has been devoted to (1). Many studies test existing analytical models and often refine them, to include, for example, random errors. There has also been some effort devoted to (2), with studies that identified new phenomenon, such as loss aversion, or regret aversion. In the future I anticipate more work devoted to testing more sophisticated operations management models.

Trend 1: Testing more sophisticated operations management models. For example, revenue management is a field ripe for laboratory investigation.

Less BOM work has so far focused on (3), testing of new institutional designs, and I expect this kind of work to be a future BOM trend. After all, operations management is by its nature a practical field, devoted to improving operations. The laboratory is ideal for cleanly and inexpensively testing supply chain mechanisms.

Trend 2: Behavioral Mechanism Design. The laboratory is ideal for better understanding how human decision-makers behave, and using this knowledge to design better systems that take into account how human decision makers are likely to behave in reality, as opposed to how they should behave in theory. Mechanisms that take human behavior into account are more likely to be implemented and to work as advertised. The laboratory is also an inexpensive way to compare alternative new designs. One example of how this approach was applied in practice is the work by Bolton, Greiner and Ockenfels (2011).

The next trend is one that I would like to see BOM researchers to consciously pursue.

Trend 3: Become more sophisticated about proposing new explanations for observed phenomena. Behavior usually deviates from predictions of standard neo-classical models, and often it does so in systematic ways. Currently the trend is to insist on full explanations, and the expectation is that a single paper should both, identify and explain a new phenomenon. But in fact this kind of knowledge and insights should be acquired in a sequence of papers, not in a single paper. Because whether an explanation is a valid one should not be based on whether it seems plausible, and even not on whether a model fits better for one explanation than for another. Instead, experiments should be designed to directly test explanations, and when appropriate, compare them directly. This can only be done through a sequence of experiments, but such an approach requires a more nuanced understanding of what it really means to "explain" a behavioral regularity.

Trend 4: More systematic investigation of the differences between students and professionals. Operations management researchers as a group seem to be more concerned with the questions regarding the effects related to the subject pool than are some other social scientists (economists, psychologists) that generally accept undergraduate students as perfectly acceptable subjects. Partly this skepticism on the part of OM researchers may have

to do with the fact that business decisions are usually made by trained managers. While other social scientists are interested in more basic research questions. I anticipate that more systematic studies of the subject pool effect are going to become trends in the future.

So far most of the studies that looked into this question failed to find any differences. While many non-experimentalists have a very strong intuition that the subject pool matters a great deal (specifically, that the student subjects are less informative that managers would be). Rather than having hypothetical arguments, I suggest that the profession should undertake systematic studies to understand in which domains the subject pool makes a difference.

Trend 5: Cultural difference. Most supply chains are multi-national, but very few laboratory experiments systematically examine cultural differences (Roth et al. 1991 is an example of such a study in economics). For our behavioral insights to be useful, we need to better understand which ones hold across cultures, which ones differ, and why.

Acknowledgement

I am grateful to Andrew Davis and Bernie Quiroga for helping proofread this manuscript and for helpful comments. Any remaining errors are my own.

References

[1] M. Allais (1953) Le comportement de l'homme rationnel devant le risqe: Critique des postulats et axioms de l'ecole americane, Econometrica 21, 503–46

[2] A.C. Atkinson and A.N. Donev (1992). Optimum experimental designs. Oxford, England: Clarendon Press.

[3] R. Axelrod (1984). The Evolution of Cooperation. Basic Books, A Member of Perseus Books Group.

[4] S.B. Ball and P. Cech (1996). Subject pool choice and treatment effects in economic laboratory research, Experimental Economics 6, 239–292.

[5] S. Blount (1995). When social outcomes aren't fair: the effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63, 131–144.

[6] G. Bolton (1991). A Comparative Model of Bargaining: Theory and Evidence, American Economic Review 8(5), 1096–1136, December.

[7] G. Bolton and E. Katok (2008). Learning-by-doing in the newsvendor problem: A laboratory investigation, Manufacturing and Service Operations Management 10(3), 519–538.

[8] G. Bolton and A. Ockenfels (2000). A theory of equity, reciprocity, and competition, American Economics Review 90(1), 166–193.

[9] G. Bolton, B. Greiner, and A. Ockenfels (2011). Engineering Trust - Reciprocity in the Production of Reputation Information, Working paper, University of Cologne, Germany.

[10] G. Bolton, A. Ockenfels, and U. Thonemann (2008). Managers and Students as Newsvendors: How Out-of-Task Experience Matters, Working paper, University of Cologne, Germany.

[11] C. Camerer (1995). Individual decision Making, in The Handbook of Experimental Economics, Volume 1 J.H. Kagel and A.E. Roth, editors, Princeton University Press, pp. 587–704.

[12] O. Carare and M.H. Rothkopf (2005). Slow Dutch Auctions, Management Science 51(3), 365–373.

[13] T.N. Cason (1995). An Experimental Investigation of the Seller Incentives in the EPA's Emission Trading Auction, American Economic Review, September 1995, 85(4), 905–922.

[14] E.H. Chamberlin (1948). An experimental imperfect market, Journal of Political Economy 56(2), 95–108.

[15] J.C. Cox, B. Roberson, and V.L. Smith (1982). Theory and Behavior of Single Object Auctions, in V.L. Smith, ed., Research in Experimental Economics. Greenwich, CT: JAI Press, 1982, 1–43.

[16] J.C. Cox, V.L. Smith, and J.M. Walker (1988). Theory and individual behavior of first-price auctions, Journal of Risk and Uncertainty 1(1), 61–99.

[17] D. Croson, R. Croson, and Y. Ren (2009). How to manage an over confident newsvendor. Working paper, Cox School of Business, Southern Methodist University, USA.

[18] T.H. Cui, J.S. Raju, and Z.J. Zhang (2007). Fairness and channel coordination, Management Science 53(8), 1303–1314.

[19] D.D. Davis and C.A. Holt (1993). Experimental Economics, Princeton University Press, Princeton.

[20] R. Engelbrecht-Wiggans and E. Katok (2008). Regret and feedback information in first-price sealed-bid auctions, Management Science 54(3), 808–819.

[21] R. Engelbrecht-Wiggans and E. Katok (2007). Regret in Auctions: Theory and Evidence, Economic Theory 33, 81–101.

[22] E. Fehr and K.M. Schmidt (1999). A theory of fairness, competition and cooperation, Quarterly Journal of Economics 114(3), 817–868.

[23] E. Filiz-Ozbay and E.Y. Ozbay (2007). Auctions with anticipated regret: Theory and experiment, American Economic Review 97(4), 1407–1418.

[24] U. Fischbacher (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments, Experimental Economics 10(2), 171–178.

[25] R.A. Fisher (1935). The Design of Experiments. Edinburgh, Scotland: Oliver and Boyd.

[26] M.M. Flood (1958). Some experimental Games, Management Science 5, 5–26.

[27] R. Forsythe, S. Horowitz, and M. Sefton (1994). Fairness in Simple Bargaining Experiments, Games and Economic Behavior 6(3), 347–369.

[28] D. Friedman and S. Sunder (1994). Experimental Methods A Primer for Economists, Cambridge University Press.

[29] J.K. Goeree and C.A. Halt (2010). Hierarchical Package Bidding: A Paper & Pencil Combinatorial Auction, Games and Economic Behavior 70(1), 146-169.

[30] J.K. Goeree, C.A. Halt, and T.R. Palfrey (2002). Quantal Response Equilibrium and Overbidding in Private-Value Auctions, Journal of Economic Theory 104, 247–272.

[31] W. Güth, R. Schmittberger, and B. Schwarze (1982). An experimental analysis of ultimatum bargaining Journal of Economic Behavior & Organization 3(4), 367–388

[32] C.A. Holt (1980). Competitive bidding for contracts under alternative auction procedures, Journal of Political Economy 88(3), 435–445.

[33] C.A. Holt (1995). Industrial organization: A survey of laboratory results, in *Handbook of Experimental Economics*, J. Kagel and A. Roth, eds., Princeton University Press, Princeton, N.J., 349-443.

[34] C.A. Holt and S.K. Laury (2002). Risk Aversion and Incentive Effects, American Economic Review 92(5), 1644–1655.

[35] R.M. Isaac and D. James (2000) Just Who Are You Calling Risk Averse? Journal of Risk and Uncertainty 20, 177–187.

[36] J. Jamison, D. Karlan, and L. Schechter (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.

[37] J.H. Kagel (1995). Auctions: A Survey of Experimental Research. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics*, Princeton University Press; Princeton, NJ. 501–585.

[38] J.H. Kagel, R.M. Harstad, and D. Levin (1987). Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study, Econometrica, Econometric Society 55(6), 1275–1304.

[39] J. Kagel and D. Levin (1986). The Winners Curse and Public Information in Common Value Auctions, American Economic Review 76(5), 894–920.

[40] J.H. Kagel and D. Levin (1993). Independent Private Value Auctions: Bidder Behavior in First-, Second- and Third-Price Auctions with Varying Numbers of Bidders, Economic Journal, Royal Economic Society 103(419), 868–79.

[41] J.H. Kagel and D. Levin (2008). Auctions: A Survey of Experimental Research, 1995 – 2008, in The *Handbook of Experimental Economics*, Volume 2, J.H. Kagel and A.E. Roth, editors, Princeton University Press, in preparation URL: http://www.econ.ohio-state.edu/kagel/Auctions_Handbook_vol2.pdf

[42] D. Kahneman and A. Tversky (1979). Prospect theory: An analysis of decision under risk, Econometrica 47, 263–291.

[43] E. Katok and A.M. Kwasnica (2008). Time is money: The effect of clock speed on sellers revenue in dutch auctions, Experimental Economics 11(4), 344–357.

[44] E. Katok and E. Siemsen (2011). The Influence of Career Concerns on Task Choice: Experimental Evidence, Management Science 57(6), 1042-1054.

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

[46] V. Krishna (2002). Auction Theory, 1st ed., San Diego, CA: Academic Press.

[47] J.A. List, S. Sadoff, and M. Wagner (2010). So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design, Experimental Economics, forthcoming.

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

[49] D. Lucking-Reiley (1999). Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet, American Economic Review 89(5), 1063–79.

[50] M. Machina (1987). Choice Under Uncertainty: problems solved and unsolved, Journal of Economic Perspectives, 1(1), 121–154.

[51] R.D. McKelvey and T.R. Palfrey (1995). Quantal Response Equilibria for Normal Form Games, Games and Economic Behavior 10, 6–38.

[52] B.B. Moritz, A.V. Hill, and K. Donohue (2008). Cognition and individual difference in the newsvendor problem: behavior under dual process theory. Working paper, University of Minnesota.

[53] T. Neugebauer and R. Selten (2006). Individual Behavior of First-Price Auctions: the Importance of Information Feedback in Computerized Experimental Markets, Games and Economic Behavior 54, 183–204.

[54] J. Ochs and A.E. Roth (1989). An Experimental Study of Sequential Bargaining, American Economic Review 79, 355–384.

[55] A. Ockenfels and R. Selten (2005). Impulse Balance Equilibrium and Feedback in First Price Auctions, Games and Economic Behavior 51, 155–179.

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

[57] C. Plott (1987). Dimensions of parallelism: Some policy applications of experimental methods, in A. Roth, ed., *Experimental Economics: Six Points of View*, Cambridge University Press, New York, NY.

[58] A.E. Roth, V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and Market Behavior, in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An Experimental Study, The American Economic Review 81(5), 1068–1095.

[59] A.E. Roth (1984). The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory, Journal of Political Economy 92, 991–1016

[60] A.E. Roth (1995a). Introduction to Experimental Economics, in The Handbook of Experimental Economics, Volume 1, J.H. Kagel and A.E. Roth, editors, Princeton University Press, 3–109.

[61] A.E. Roth (1995b). Bargaining Experiments, in The Handbook of Experimental Economics, Volume 1 J.H. Kagel and A.E. Roth, editors, Princeton University Press, 253–248.

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

[63] A. Rubinstein (1982). Perfect Equilibrium in a Bargaining Model, Econometrica 50(1), 97–109.

[64] A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom, and J.D. Cohen (2003). The neural basis of economic decision-making in the ultimatum game, Science 300, 1755–1758.

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

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

[67] V.L. Smith (1962). An Experimental Study of Competitive Market Behavior, The Journal of Political Economy 70(2) (Apr., 1962), 111–137.

[68] V.L. Smith (1976). Experimental Economics: induced Value Theory, American Economic Review 66(2), 274–279.

[69] V.L. Smith (1982). Microeconomic Systems as an Experimental Science, American Economic Review 72, 923–955.

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

[71] L.L. Thurstone (1931). The indifference function, Journal of Social Psychology (2), 139–67.

[72] W. Vickrey (1961). Counterspeculation, auctions and competitive sealed tenders, The Journal of Finance 16(1), 8–37.

[73] L. Von Neumann and O. Morgenstern (1944). Theory of games and economic behavior, Princeton University Press, Princeton, NJ.

[74] W.A. Wallis and M. Friedman (1942). The empirical derivation of indifference functions, in Studies in mathematical economics and econometrics in memory of Henry Schultz, O. Lange, F. McIntyre, and T.O. Yntema, editors, Chicago University Press, Chicago, IL 175–89.

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

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