1. Introduction
2. The experimental method & mindset
3. Methods to establish rational valuation
   3.1 The divergence between WTP & WTA
   3.2 Preference Reversals
   3.3 Synopsis on rational valuation
4. Methods to measure values
   4.1 Experiments valuing public goods
   4.2 Experiments valuing private risks
   4.3 Synopsis on direct measurement
5. Methods to testbed surveys
   5.1 Hypothetical bias I: Does it exist?
   5.2 Hypothetical bias II: Calibration
   5.3 Framing I: Surrogate bidding (scope/embedding)
   5.4 Framing II: Incentive Compatibility
   5.5 Synopsis on testbedding
6. Concluding remarks
Abstract
This chapter explores how economists use experimental methods to understand better the behavioral underpinnings of environmental valuation. Economic experiments, in the lab or field, are an attractive tool to address intricate incentive and contextual questions that arise in assessing values through direct statements of preferences. By combining empirical observation with theoretical insight, researchers use the experimental method and mindset to help explain how economic and social contexts matter to valuation. Herein we consider three themes in applying the experimental method to valuation—rational choice theory and stated values, direct value elicitation in the field and lab, and “testbedding” survey designs prior to field application. First, experimental tests of rational valuation are discussed. This lab work examines whether respondents make choices and state values in a manner consistent with standard rational choice theory. The circumstances of rational valuation are illustrated by the malleability of two classic anomalies—the WTP-WTA divergence and the preference reversal phenomenon. Second, direct experimental methods to measure actual values for public and private goods are examined. These experiments ask people to buy and sell actual goods to elicit real values, in which researchers test how alternative exchange institutions affect these values. Third, we survey testbed experiments designed to identify potential incentive problems caused by hypothetical valuation questions. Four topics are discussed: testing for hypothetical bias, calibrating real and hypothetical values, examining surrogate values (or scoping) for specific environmental preferences, and evaluating the incentive (in)compatibility of alternative elicitation mechanisms.

JEL classification codes (B4, C9, H4, Q2)
1. Introduction

How do people value environmental protection? This question has captivated economists for decades. Driven by academic curiosity, Presidential Executive Orders, and court cases over natural resource damage assessment, the United States has witnessed a push to produce absolute numbers so that society can add nonmarket considerations onto the benefit-cost balance sheets that help shape government policymaking and legal decisions. The same is also true of many other nations, e.g., Sweden, United Kingdom. It might not be an overstatement to say that probably as much collective energy has been spent on defining, estimating, and debating nonmarket valuation as any other topic in environmental economics. And in particular, people have pondered the challenge of control and accuracy in nonmarket valuation through public opinion surveys, i.e., contingent valuation and other stated preference methods.

The desire for more control and accuracy in revealing nonmarket values through direct questions about actual goods led researchers like Peter Bohm, Richard Bishop, and Thomas Heberlein in the 1970s toward the methods of experimental economics and the insights gained from the experimental mindset that emerged in the 1960s. While Bohm, Bishop and Heberlein were primarily interested in field valuation work, a typical economics experiment is run in a laboratory just like in chemistry or physics. Here test tubes and particle accelerators are replaced with experimental instructions and networked computers programmed to create interactive markets, auctions, and strategic games. Monitors then observe the actual behavior of people within these institutions given initial endowments of rules, resources, technology, and preferences. Over 100 experimental economics labs now exist around the world, to which experimentalists recruit both students and non-students to participate to test theory, look for patterns, and to explore new phenomenon that arise in choices made within and outside of exchange institutions.

The push to make the experimental method a part of nonmarket valuation began in earnest with the state of the art assessment of contingent valuation summarized in Cummings et al., (1986), and the early work of Knetsch and Sinden (1984), Coursey and Schulze (1986), and Bennett (1983). Today, more researchers draw on the experimental mindset, in both the lab and the field, to understand better how people learn about incentive systems and what this might imply for valuing environmental protection.

This chapter explores how economists use the experimental method to better understand the behavioral underpinnings of environmental valuation. We consider three intertwined areas of the experimental method in valuation: (i) methods to test rational choice theory, the principles that define the economic surplus measures upon which nonmarket valuation rests; (ii) direct methods to measure actual values for public and private goods, in the field and the lab; and (iii) methods to use the lab as a “testbed” for defining incentive systems prior to the field application of hypothetical surveys. As with any survey, the chapter focuses its attention on a few themes. Other reviews and opinions on experimental methods in environmental economics include Cummings et al. (2000), Cummings and Taylor (2001), Harrison (1996), Shogren and Nowell (1992), Shogren and Hurley (1999), and Shogren (1993, 2001).
2. The Experimental Method & Mindset

With all the available theoretical and empirical tools in economics, why bother do an experiment? Three reasons come to mind—to test theory, for pattern recognition, and for testbedding. First, one can use experiments to test a priori expectations about behavior, rational or otherwise. Researchers use experiments to test the predictive power of a theory, to test the robustness of the axioms underlying the theory, to test the specific boundaries of a general theory, and to measure the gradient of behavioral change (i.e., comparative statics). Experiments provide a sterile environment to test theory by controlling for noise and other confounding factors. Given this control, a researcher can assess the ability of a theory and its alternatives to organize observed behavior. For valuation work, the lab is used to test whether stated values are consistent with economic theory (e.g., the divergence between willingness to pay and willingness to accept measures of value).

Second, people use the lab to look for patterns of behavior. This more pragmatic use of the lab allows people to explore how people construct preferences and beliefs, identify and measure breakdowns from rationality, examine how contextual frames affect behavior, determine reactions to new information, and consider how people coordinate actions voluntarily and under duress. Pattern recognition can provide the motivation for people to develop alternative theories (i.e., prospect theory) based on ex post explanations of observed behavior. Howard Raiffa’s (1982) work on negotiation is a good example of pattern recognition. Raiffa created a classroom “quasi-laboratory” in which students negotiated over alternative controlled scenarios. They then interpreted their observed behavior, so as to design modified experiments and resample new subjects. They then collectively discussed whether the heuristic insight gained from observing actual behavior in the lab could translate into real world applications. For valuation work, pattern recognition involves the direct elicitation of values for goods or services in a lab auction or field exchange given alternative incentive mechanisms, endowments, information sets, and with and without the signals set through repeated market experience.

Third, laboratory experiments are used as a testbed for economic design—the construction of new institutions and mechanisms designed for efficient resource allocation. The most prominent example is the use of the lab to test pilot the efficiency of the proposed FCC spectrum auctions (see for example Plott, 1994), to compare alternative policy options, to explore how friction affects efficiency and the distribution of wealth, and to consider how institutional power can transform patterns of behavior. An example of valuation testbedding is examination of the potential incentive compatibility of alternative value elicitation mechanisms (e.g., discrete choice or referendum questions). The ultimate success of testbedding different mechanisms rests in the open question of external validity—is the behavior in the lab a reliable guide to behavior in the real world? Experimental economists believe so, albeit within reasonable limits. Lab results represent real evidence about how certain people will behave in a given economic environment. Additional real-world complexity can then be added into the lab environment in a controlled fashion to identify likely conditions that might cause a mechanism to fail in the wilds.

The wide-ranging idea that experimental research provides the definitive evidence prompted worldly philosophers like Bishop George Berkeley and John Stuart Mill to take
the extreme position that "...nothing beyond experimental knowledge is either possible or necessary" (Kline, 1985, p. 19). This predilection toward experiments as science led ecologist Heath (1976, p. 3) to claim: “Science consists essentially in an attempt to understand relations of selected aspects of things and events in the real world, an attempt which should have both intuitive and logical components, and which must be based on observation and tested by further observation. This definition of course excludes mathematics, which does not have to be based on observation (data) but only on postulates which need not have any relevance to the real world.”

While more inclined to blend the empiricist and rationalist traditions, John Herschel’s (1997[1830]) 19th-century discourse on the philosophy of science still elevated experiment as method: “By putting in action cases and agents over which we have control, and purposely varying their combinations, and noticing what effects take place; this is experiment…. We cross-examine our witness, and by comparing one part of his evidence with the other, while he is yet before us, and reasoning upon it in his presence, are enabled to put pointed and searching questions, the answers to which may at once enable use to make up our mind. Accordingly it has been found invariably, that in those departments of physics where the phenomena are beyond our control, or into which experimental enquiry, from other causes, has not been carried, the progress of knowledge has been slow, uncertain, and irregular; while in such as admit of experiment, and in which mankind have agreed to its adoption, it has been rapid, sure, and steady” (pp. 76-77).

Obviously others question such passionate beliefs on experimentation. Theory and non-experimental empirics have long dominated the economist’s toolbox, causing some to ask whether economists who use the small-scale experimental method and mindset have anything to say about large-scale environmental risks in the wild. Many economists who run experiments, including myself, believe the answer is yes—the experimental method matters for environmental economics. Experiments backup with data the idea that economic incentives matter, and that by addressing such behavior explicitly in the lab we can better understand the ultimate success and failure of a policy. While no general panacea, experiments have many uses in the policy arena: to support or counter some specific policy initiative; to increase the costs to policy makers who choose to ignore economic behavior; to shift the burden of proof in a policy debate; to demonstrate how friction (or the lack of it) affects a policy; to reveal how (in)sensitive benefits & costs are to context and frames; to reveal how people tend to react to absolute and relative levels of risk; and to support the notion of why and when flexibility can lead to more environmental protection at lower cost. The challenge is to not oversell the lab results.

The defense of the experimental method in general, and applied to environmental issues in particular, rests on the foundations laid down by Vernon Smith, Charles Plott, and Peter Bohm. Their path-breaking work and the ensuing literature has moved economics further toward being considered an experimental science, with its own set of protocols and rules for the lab and field (see the chapters in Kagel and Roth, 1995). Today economists who use experiments to address environmental policy have a lot in common with natural scientists. Although economics has traditionally devoted its energy toward theory and empirical work based on field data, environmental economists now use controlled experimentation like their colleagues in the life sciences (e.g., biology and
ecology), the people upon whom environmentalists usually rely on to make the case for more environmental protection (Shogren and Nowell, 1992; Shogren, 2001).

Figure 1 illustrates a basic triad that illustrates the experimental method and mindset. The triad reflects the three components that underlie an experiment: the initial endowment that defines the human and natural environment, \(E\); the institution or mechanism of exchange \(I\); and the actual behavior of the subjects \(A\). The environment includes basic economic endowments like preferences, technology, physical constraints, property rights, and information structure. The institution specifies the rules that aggregate and generate information and coordinate actions, and it outlines the rules of exchange and its consequences. Repeated exposure to the institution is common practice in the lab so that people have the opportunity to gain experience with the institution, new information, and their own mind to better understand their endowed or personal preferences. Given the environment, people send a message \(B\), which could be an auction bid or stated value to the institution. Based on the set of messages received, the institution then allocates resources and costs, \(G\), given the known rules of exchange.

Researchers then observe how people actually make choices in the lab or field, and how this behavior matches up with a specific performance criterion like Pareto efficiency \(P\).

Based on this triad, we categorize environmental economic experiments into two groups—institutional and valuation. Institutional experiments control the environment \(E\) to explore how alternative market and nonmarket mechanisms \(I\) affect the allocation of scarce resources. Evolving from informal games and role-playing, these experiments are now used to address the question of economic design, i.e., the efficient institutional design given administrative failure, public goods, externalities, asymmetric information, and incomplete markets. By far, institutional experiments dominate the experimental economic literature. In a later chapter in this Handbook, Peter Bohm considers institutional experiments examining control issues in environmental policy.

In contrast, valuation experiments stand the experimental triad on its head (Coursey, 1987). Now the researcher wants to control the institution and actual behavior through a design that generates predictable patterns of behavior. If a researcher wants people to sincerely reveal their preferences about a good or service, they can employ an exchange mechanism that is demand-revealing in theory, e.g., the uniform-price auction. If the goal is to create an upper and lower bound on potential values, one can design the lab environment to create incentives for people to either overstate their values or understate their preferences for the good.

The key is that the lab is flexible, the design is under the researcher’s control, and the decisions are real. Real choices, albeit stylized, are another key to using the experimental method in valuation work. The common belief is that performance-based payoff systems produce behavior that more closely parallels real life or market behavior than flat-rate payoff systems or no payoff at all. Most economists are suspicious of hypothetical choices because by definition these decisions are not salient—people do not have an unqualified right to the corresponding rewards or punishment because they are hypothetical. Saliency is one of Smith’s (1982) five sufficient conditions for a controlled economics experiment. Bergstrom and Stoll (1989) make the case that hypothetical choices are indeed salient if people get benefits from simply making a hypothetical choice; if they are altruists; or if they expect some reward based on their choice to come to them at some date in the future. While creative, it stretches things to argue that people
always believe their answers will be translated in full to social policy, either today or in
the future. It is also a stretch to presume that goodwill and altruism are one-to-one
proxies for the rational choices people are presumed to be making about a hypothetical
good.

The issue of saliency reveals that at its core, the experimental method is about
control—the researcher wants to control institutions, endowments, incentives, and
sometimes preferences, and then observe motivated economic behavior. This holds in
general, not just for valuation work (see Harrison, 1992). An experimenter should try to
control for all the confounding factors in the design by holding \( n-1 \) factors constant; and
he should make the rewards significant enough so subjects are motivated to make rational
decisions. The goal is to design experiments to avoid the classic problem of under-
identification, in which two or more variables change at once, either in the lab’s external
environment or in the person’s internal environment (e.g., preferences for more wealth
versus preferences to be viewed as generous) so as to confound one’s understanding of
cause-and-effect. The loss of control implies that a researcher does not know whether
unpredicted behavior is due to a weak theory or a poor experimental design, or both.\(^6\)

But one researcher’s gain of control can be another’s loss of context. These
researchers argue that it is these confounding factors (e.g., uncontrolled body language
sent in face-to-face bargaining) that provide the rich economic context that motivates
real-world behavior. They point out that striving for complete control is self-defeating
because it creates an economic environment that is too sterile, too abstract, too unreal,
with no parallel in the real-world, and therefore generates subjects who are motivated by
salient payoffs, but unmotivated by the context in which these payoffs are earned. The
inspiration is that economic theory should be robust enough to explain actions in real-
world social interactions that have copious contextual cues that occur simultaneously. In
response, advocates of more control point out that too much context makes the
experiment too much like an open-ended case study or descriptive naturalist observations.
Experiments based on too many confounding factors and misidentified motivation yields
no universal patterns based on first principles, which creates little ability to generalize
beyond the specific case. One might even argue this approach is non-experimental. The
recommendation is that when examining such contextual studies one should note whether
the method permits the conclusions that the author has made. A causal inference may
have been made when the design does not allow such inference.

But this debate is not new. Fifty years ago, Harvard chemist James Conant (1951;
p.56) defined this deep-rooted conflict of control and context within a scientific discipline
as the degree of empiricism—“the extent to which our knowledge can be expressed in
terms of broad conceptual schemes.” His notion characterized a discipline as falling
along a continuum determined by the fraction of work in controlled hypothesis testing
and contextual pattern recognition. Conant expands on the confluence of scientific
experimentation and the “sooty empiric”:

"About three centuries ago the trial-and-error experimentation of the
artisan was wedded to the deductive method of reasoning of the
mathematician; the progeny of this union have returned after many
generations to assist the ‘sooty empiric’ in his labors. In so doing the
applied scientist finds himself face to face with one of his distant
ancestors, so to speak….In those practical arts where the degree of
empiricism is still high, men with the most advanced scientific training and using the latest equipment will often have to resort to wholly empirical procedures. On the one hand they will labor to reduce the degree of empiricism as best they can; on the other they must improve the art by using the knowledge and methods then at hand. In short, advances in science and progress in the practical arts today go hand in hand” (p.59-60).

Conant’s remarks hold today for stated preference methods in valuation. Here the degree of empiricism is high. Nothing comparable to the theory of market efficiency exists in stated preference work that provides a mathematical basis for how a person’s hypothetical intentions toward the environment translate into his actual actions (see section 5.2). Hoehn and Randall’s (1987) theory of value formation is the rare attempt to develop a theory that relates stated preference answers to valuation questions. They maintain that respondents are still rational and motivated but have imperfect information and incomplete optimization about the good in question. But the economics of hypothetical or unmotivated behavior has not been formulated into broad theoretical terms of general acceptance. That is why observing behavior under controlled conditions in the lab still matters. But we are dealing with people not protons, so there are degrees of control and the relevant decision is rarely either-or, but how much. We also need to understand why and how people react in the real-world environmental entanglement of markets, missing markets, and no markets.

The lab is the natural milieu to help reduce the degree of empiricism in this search. And the debate over control and context will continue on. What is unique to one research can be bizarre to another. The challenge is to find the “correct” balance of control and context to create an experimental design for an economic situation in which the experimenter is confident that people are motivated to do what they do for the reason she thinks they are doing it. Searching for this correct balance arises because environmental economic experiments, especially valuation studies, lie at the intersection of three general areas of experimental research—traditional experimental economics that explicitly considers markets, standard experimental psychology that presume no markets, and environmental economics that acknowledges that markets are missing (see Figure 2).

Consider each area. Experimental economics usually examines the behavior of people within the context of an active exchange institution like the market. The market allows a rational person to exploit non-rational choices, which then helps to reinforce rational decision-making throughout the population. Psychological experiments usually focus on isolated individual behavior, and decisions made outside an exchange institution. Isolated choices are then compared to a benchmark model of rational choice (e.g., expected utility theory). Valuation experiments must confront the issue that people make decisions in both market and nonmarket settings, and that this interaction of exchange institutions could well affect the values that are elicited. This intersection of methods requires the researcher to explore all three literatures because they are intertwined in valuation work. The lab can be very helpful in pulling together and sorting out isolated and socialized choices and statements of value. The goal is to better understand how people make choices and form values given they make decisions both inside and outside of markets.
This recognition that valuation methods need input from many sources has forced many people interested in valuation to be quite pragmatic in their research style. They look for patterns, either deviations from standard rational choice theory or new, unexpected behavior that emerges, similar to experimentation in the life sciences (see for example Heath, 1976; Hölldobler, 1985). Pragmatism implies that methods and choices result from the workability of common sense rather than formal, predetermined rules of evidence. Questions of method are answered by experience about what works rather than on preconceived notions of methodological principles. And those economists who use the lab to study environmental economic phenomena learn to appreciate our ancestors who lived when pragmatism ruled. Limited capacity, time, and money imposed a discipline on those forerunners to identify and test the critical valuation questions.

Experimental economics demands the same discipline today. One identifies a phenomenon, designs an experiment, runs the treatments, and explores the restrictions that best organize observed behavior. Working along with theory and prior empirical information, experiments become habitual; another productive approach to discipline one’s thinking about economics and valuation. This chapter focuses on these valuation experiments. We consider how the experimental method has been used in valuation research over the past three decades, starting with methods to establish whether statements of economic value are rational as defined by traditional welfare economics.

3. Methods to establish rational valuation

Rational choice theory and welfare economics provide the traditional analytical foundation upon which conventional environmental valuation methods rest (see Mäler, 1985; Freeman, 1993). The presumption of rational choices made within active exchange institutions underpins the economic definition of value (Samuelson, 1938). People are assumed to have core preferences that reflect what they want, which then they can articulate into monetary values. If a person’s preferences satisfy well-defined axioms, his behavior can be modeled as if he is maximizing his well-being. One’s stated preferences for changes in the level of environmental goods and services have purpose and meaning because they are grounded by a core set of preferences and a binding budget constraint.

But researchers have used laboratory experiments to challenge rational choice theory, and hence economic approaches to value the environment. Psychologists and economists have evidence of systematic deviations in behavior from the predictions of rational choice. Their experiments have revealed that behavioral anomalies abound when people confront risks similar to those emerging in environmental issues, like climate change. Classic examples include inconsistent preferences when choosing between risky and safer gambles (Allais 1953); extra aversion to risks that have ambiguous probabilities (Ellsberg 1961); and a systematic discrepancy between choosing and valuing alternative gambles, i.e., the preference reversal phenomenon (Lichtenstein and Slovic 1971). These anomalies, and many others, matter for the rational valuation of environmental protection because they undercut the internal validity of the cost-benefit estimates economists use to evaluate alternate protection strategies (see Machina, 1987; Thaler 1992; Camerer, 1995; Rabin, 1998).

Challenges to rational choice theory are threats to rational nonmarket valuation. If people do not follow rational choice theory, and instead make irrational or inconsistent
choices or if values are momentary declarations, one becomes concerned that preferences and stated values are transient artifacts of context (see Tversky and Simonson, 1993). And if uncontrolled or unmeasured contextual cues affect stated values, the challenge is to judge whether stated values represent some notion of “true” preferences. If stated values do not reflect preferences, economic behavior is unlikely to have the welfare-economic implications normally ascribed to choices and valuations. This threatens the validity of the cost-benefit estimates economists use to evaluate alternate policies, regardless of whether one is using experimental methods, market prices, or hypothetical surveys.

This section uses two examples to illustrate how the experimental method can be used to assess whether valuation is rational—the disparity between the willingness to pay (WTP) and the willingness to accept (WTA) compensation measures of economic value, and the case of preference reversals. These two examples illustrate the challenge in establishing whether valuation is based on rational choice theory.

3.1 The divergence between WTP and WTA

The experimental method has played a prominent role in addressing one of the most intriguing questions of rational valuation—the gap between WTP and WTA measures of value. Rational choice theory suggests that with small income effects and many available substitutes, the willingness to pay for a commodity and the willingness to accept compensation to sell the same commodity should be about equal (see Hanemann, 1991). But evidence has accumulated over the past two decades that suggests that a significant gap exists between WTP and WTA. WTA exceeds WTP by up to tenfold.\(^{10}\)

Since these value measures are used to help guide public policy decisions, the divergence raises questions about which measure to use in actual practice. If the decision is to conserve some environmental amenity, using a WTA measure could likely generate a significantly greater economic value for conservation than would a WTP measure—perhaps enough of a difference to tip the balance towards conservation. But if WTA is based on irrational or inconsistent behavior, its usefulness for policy can be questioned.

The first lab experiment to examine the WTP-WTA divergence was Knetsch and Sinden (1984). They asked subjects to value lottery tickets, in which the prize and price levels were selected to avoid wealth effects. The subjects were split into two groups—half were given a lottery ticket and then offered cash to sell the ticket. The other half was offered the chance to buy a ticket and play the lottery. Using identical money amounts, the percentage of people willing to pay differed significantly from the percentage of people willing to accept. This was taken as evidence that a gap existed between WTP and WTA. When participants were quizzed on how they would advise another person who faced the same situation, no significant difference was observed.

Coursey et al. (1987) reconsidered the WTP-WTA question. They asked whether experience with an incentive compatible auction might remove the gap. While far from the acme of perfection, their experiment deserves attention because it was the first test of the WTP-WTA gap within an active lab exchange institution. Their challenge was to assess whether rational valuation should be defined as an individual or a social construct. Psychologists usually view rationality as an individual construct, whereas most economists view it as a social construct, in which active exchange institutions work to reinforce rational decisions.\(^{11}\) The advantage of laboratory experimentation over survey
measures of WTP and WTA is that with repeated trials the subjects can learn about the market and the market-clearing price.

Students at the University of Wyoming were told about the ‘very bitter-tasting and unpleasant substance,’ sucrose octa-acetate (SOA), and the asked hypothetical questions about WTP and WTA values. Students then tasted a drop of the SOA and were asked to bid again. Monitors then tried to lower or raise the stated bids to determine what they called a semi-hypothetical iterative bid. Finally, the design used a modified version of the uniform price Vickrey (1961) auction to elicit the WTP and WTA values: a fifth-price auction with a veto rule. Here the four highest-bidders each bought one unit of a good and paid the fifth highest bid, provided none of the four winners vetoed the exchange.

In general, the Vickrey auction is incentive compatible, provided there is no veto rule. The auction is designed to induce people to sincerely reveal their private preferences for new goods and services. People have an incentive to tell the truth because the auction separates what they say from what they pay. Sincere bidding is the weakly dominant strategy. Underbidding risks foregoing a profitable purchase, whereas overbidding risks making an unprofitable purchase. Evidence from induced value experiments suggests the auction can produce efficient outcomes in the aggregate, although the average person tends to overbid (see Kagel, 1995). The appeal of the auction for valuation work is that it is demand revealing in theory, relatively simple to explain, and it has an endogenous market-clearing price.

The veto rule changes the incentives of the auction, however, because no bidders are connected through the veto. Coursey et al.’s auction, therefore, is no longer incentive compatible, so the following results must be assessed with this knowledge. Their results suggest that while hypothetical bids displayed a large disparity, mean WTP and WTA bids converged with repeated exposure to the Vickrey auction environment. The convergence of the bids with market experience was argued to support received theory of rational choice—preferences were independent of initial endowments. This rational value convergence story, however, was weakened significantly by the observation that one outlier had a substantial effect on the average values (see Gregory and Furby, 1987).

Knetsch and Sinden’s reply to Coursey et al. revealed little deference for the iterative market explanation, and in fact hinted at the potential flaw with Vickrey with veto auction (see Harrison, 1992). Knetsch then came back with a series of new experiments aimed at revealing why the divergence is a reality and not some transient artifact. First, Knetsch (1989) attributes the willingness to pay and willingness to accept to the endowment effect—people are less willing to surrender something they own compared to their eagerness to acquire something new. One of several experiments conducted by Knetsch consisted of giving half of the participants a candy bar and the other half of the subjects a coffee mug of approximately the same value. Subjects were then offered the opportunity to trade for the other commodity. Preferences for the mug over the candy bar varied from ten to eighty-nine percent depending purely on which commodity the person was given first. Other experiments by Knetsch offered similar results (see for example Borges and Knetsch, 1998).

The next influential paper was Kahneman, Knetsch, and Thaler (1990), who report experimental evidence to support their idea of the endowment effect as why WTP
diverges from WTA. The effect exists if people offer to sell a commonly available good in their possession at a substantially higher rate than they will pay for the identical good not in their possession (e.g., pens and coffee mugs). To illustrate the robustness of the effect, they used a Becker-DeGroot-Marschak (1964) mechanism to elicit preferences in the fifth treatment. Like the Vickrey auction, the BDM mechanism separates what people say from what they pay. A person’s weakly dominant strategy is to state her true WTP or WTA. Unlike Vickrey, the BDM market-clearing price is exogenous. In addition, the BDM is with a single individual, not with a group as in the Vickrey auction. After collecting all bidding sheets from buyers and sellers, a market-clearing price was randomly selected from some distribution known by the subjects. A buyer willing to pay at least this market price bought the commodity; a seller willing to accept less than or equal to the price sold the commodity. The market price and buyers and sellers were then determined (also see Grether and Plott, 1979).

The results from the Kahneman et al. treatments make a case for the existence of a fundamental endowment effect—WTA exceeded WTP in all treatments over all iterations. People’s preferences seemed to depend on initial endowments of resources, a violation of rational choice theory. The idea behind this psychological argument for the endowment effect is that people treat gains and losses asymmetrically—the fear of a loss weighs more heavily than the benefit from an equal gain. People who like the status quo demand relatively more to accept a change for the worse. Additional support for this view is found in MacDonald and Bowker (1994), who support the existence of an endowment effect based on the significance of perceived property rights on WTP for improved air quality versus WTA to forgo better air quality (also see Morrison, 1998).

In contrast, the lab valuation experiments of Shogren et al. (1994) observed no significant divergence between WTP and WTA for similar goods. They designed an experiment to test the proposition that given positive income elasticity and repeated market participation, WTP and WTA will converge for a market good with close substitutes (e.g., the candy bar and mugs) but will not converge for the nonmarket health risk with imperfect substitutes (risky sandwich). They also tested for whether transaction costs could explain the WTP-WTA gap by having a large supply of the market goods right outside the lab door that could be purchased by the subjects once the experiment was over. The results showed that WTP and WTA values did converge with repeated trials for candy bars and mugs—goods with many available substitutes; but that the values continued to diverge for the reduced health risks from safer food—one’s health has little substitutes.

A key difference between the two experiments is that Shogren et al. used a Vickrey second-price auction with endogenous market-clearing price feedback. These results raise the question of whether the endowment effect is a fundamental part of choice or simply an artifact of a weak exchange environment. The weaker the exchange institution, the weaker the socialization of rational behavior and the stronger the potential hold of asocial anomalies on choice. If the objective is to elicit values for commodities in market-like settings that punish mistakes and reward good decisions, an exchange institution such as the Vickrey auction with repeated trials is appropriate.

The question is why the different results in the Kahneman et al. and Shogren et al. experimental designs? Was the difference is auction enough to cause a difference in values? Earlier lab work on alternative market institutions suggests that the answer might
be “yes”. But too many differences existed across the two experimental designs, however, to confirm or reject this suspicion by just looking at the existing data. In response, Shogren et al. (2001a) address this issue by designing an experiment to specifically address this mechanism-dependence question, in which the auction was the treatment. The thesis was that if it is the endowment effect that accounts for observed behavior, the effect should be observable and persistent for any mechanism used to elicit WTP and WTA, provided the mechanism is incentive compatible. They test this thesis by evaluating the impact of three auction mechanisms in the measurement of WTP and WTA measures of value for goods with close substitutes—the Becker-DeGroot-Marschak mechanism with random, exogenous price feedback, Vickery’s second-price auction with endogenous market-clearing price feedback, and a random nth-price auction with endogenous market-clearing price feedback (see Fox et al., 1998).

The random nth-price auction combines elements of the Vickrey auction and the BDM mechanism, such that the market-clearing price is random but endogenously determined. Randomness is used to engage all bidders, and to reduce the incentive to fixate on any given price. A random nth-price auction works as follows: each bidder submits a bid (offer); each bid (offer) is rank-ordered from lowest to highest; the monitor selects a random number—the n in the nth-price auction, uniformly-distributed between 2 and n (n bidders); and in the WTP case, the monitor sells one unit of the good to each of the n-1 highest bidders at the nth-price; in the WTA case, the monitor buys one unit each from the n-1 lowest bidders and pays the nth-lowest bid.

Table 1 summarizes the design parameters in Kahneman et al. (1990), Shogren et al. (1994), and the new treatment. The experimental parameters are auctioned goods: a brand-name candy bar in Stage 1 and an University coffee mug in Stage 2; initial monetary endowment: $15 paid up-front; number of trials: ten trials per experiment, in which wealth effects were controlled by randomly selecting one trial to be binding; retail price information: none was provided; subject participation: voluntary students participants; number of subjects per session: 8-10 subjects in the second-price auctions and 20 for the BDM mechanism; and the three auction mechanisms.

The results from these new experiments show that initial bidding behavior in each auction did not contradict the idea of an endowment effect. Figure 3, for example, shows the ratio of mean WTA and WTP bids across trials and auctions for the university mug. We see that if it is an endowment effect that originally governs bidding, these results show that the effect can be eliminated with repetitions of a second-price or random nth-price auction. If the thesis is correct that an endowment effect should persist across auction mechanisms and across trials, these experimental results show that the case for such a fundamental effect is open to challenge. Rutström (1998) also observed behavioral differences across incentive-compatible auctions. She detected WTP differences elicited in the BDM and Vickery auctions.

3.2 Preference Reversals

Expected utility theory is the cornerstone of modern decision making under risk, and consequently the cost-benefit analysis of environmental protection. The theory presumes that people are fairly sophisticated such that they can evaluate both old and new gambles consistently. Laboratory evidence emerged, however, which showed that many people reverse their preferences. A preference reversal is said to occur when a
person’s choices – a direct reflection of his preferences – between two options is inconsistent with the ranking of his buying/selling prices – an indirect reflection of his preferences. The preference reversal phenomenon contradicts the presumption that elicited preferences should be invariant to the elicitation method.\textsuperscript{12}

The preference reversal phenomenon is one of the best-documented violations of rationality.\textsuperscript{13} Grether and Plott (1979), doubtful of the robustness of psychology experiments, ran their own economic experiments to explore whether economic incentives would stop this inconsistency. Their lab evidence, however, revealed a robust pattern of preference reversals, despite inducements like greater rewards, different presentations, training, and record keeping. The phenomenon has caused some observers to conclude that traditional valuation exercises and choice-based surveys are unreliable, and that economists need to channel resources to “develop new methods for valuing environmental resources” (Irwin et al., 1993). This is a serious charge because it implies that preferences for environmental protection are ephemeral, affected by poorly understood psychological contextual cues, and therefore of limited use in traditional cost-benefit analysis. If researchers who want to estimate the benefits of protection ask people their willingness to pay for lower risk, and people state values for this risk reduction that are inconsistent with their underlying preferences, researchers obtain no useful information with which to judge the benefits of alternate policies.

Consider a common example used to induces preference reversals. A monitor presents a person with some variation of the following pair of bets and asks him to choose one bet out of the pair:

\[
\text{Pbet: } p \text{ chance of } X \quad \text{q chance of } Y \\
1-p \text{ chance of } x \quad 1-q \text{ chance of } y
\]  

where \( X > x \), \( Y > y \), \( p > q \), and \( Y > X \). The subjects were then asked to value each bet by stating the maximum (minimum) they were willing to pay (accept) to buy (sell) the bet. A specific gamble commonly used is:

\[
\text{Pbet: } 35/36 \text{ chances to win } \$4 \quad \& \quad 1/36 \text{ chance to lose } \$1 \\
\text{$bet: } 11/36 \text{ chances to win } \$16 \quad \& \quad 25/36 \text{ chances to lose } \$1.50
\]  

The two gambles have the same approximate expected value, $3.86. Expected utility theory requires that the bet selected would also be the bet that was valued the highest. Usually around half the subjects from any given population violated this prediction by choosing the Pbet and assigning a higher value to the $bet.

While preference reversals have caused some theorists to develop alternative models with non-transitive preferences over bets,\textsuperscript{14} others have questioned its robustness for choices in market settings with non-trivial lotteries and arbitrage. Knez and Smith (1987), for example, challenged the robustness of the phenomenon with market trading and repeated responses. They found that in double auction asset trading, market values followed expected utility theory even if individual responses did not. Individual responses even approached rational behavior with repeated trials. In addition, Cox and Epstein (1989) found consistent preference reversals, but observed that by paying
subjects for each gamble instead of randomly choosing one, subjects reversed symmetrically. Chu and Chu (1990) showed how arbitrage eliminates preference reversals quickly; Cox and Grether (1996) observed that the phenomenon tended to disappear with monetary incentives, immediate feedback, and repetition.

Consider three market tests in detail. Bohm (1994a,b) and Bohm and Lind (1993) designed three field experiments to determine whether the robust laboratory results hold up in the field with (i) choice objects of non-trivial values, (ii) elicited bids or asks for real auctions, and (iii) engaged subjects who were engaged with the lottery. Bohm (1994a) used non-trivial real choice objects which the test subjects had revealed an interest in buying, and in which uncertainty was characterized by subjective probabilities and more than two outcomes (Bohm, 1994a). Used cars (a Volvo and an Opel) were selected as an example of ‘non-lottery’ objects whose performance (outcome) is uncertain. Two cars were bought -- both priced at SEK (Swedish kronor) 8,000 or $1,200 1990 US dollars—to be sold at an auction, offered to some 2,000 students at Stockholm University. Twenty-six showed up, test drove the cars, checked inspection reports, and participated in a hypothetical second-price auction (of two bicycles) to make them more familiar with this type of auction. Given the reputations of the two cars, the Volvo bore some resemblance to a Pbet and the Opel to a $bet. No preference reversals were observed—bids on the preferred car never fell short of bids on the other car.

Bohm and Lind (1993) challenge preference reversals by constructing an experiment that used real-world lotteries with more than two (7-12) levels of prizes and separated the subjects interested in buying lottery tickets from those who were not interested. Student subjects choose between three options: (i) SEK 40 ($7 US dollars in 1991) in cash, (ii) an SEK 40 share of a package of $bet-type Swedish state lottery tickets, with the highest prize being SEK 1,000,000; and (iii) an SEK 40 share of a package of Pbet-type Swedish state lottery tickets, which had a highest prize of SEK100,000. Both lottery options have expected payoffs of SEK 20 and at least seven prize levels. As a comparative benchmark to mimic earlier work, subjects were also asked to choose between lotteries that had similar expected values with smaller variance. Relative to the benchmark, subjects reversed preferences significantly less with real-world lotteries—23 percent versus 73 percent.

Bohm (1994b) revisited the question of whether people reverse their preference when confronted with temporal choice, as suggested by Tversky, Slovic and Kahneman (1990). In hypothetical cases, some people are inconsistent in their choice of and values for claims redeemed at different future dates. For instance, a person’s choices and values were inconsistent for a short-term claim of $1,600 received in 1.5 years versus a long-term claim of $2,500 received in 5 years. Short-term claims and long-term claims play a role similar to that of Pbets and $bets. Bohm tested the robustness of this hypothetical finding using real claims and subjects expected to be confronted with decisions of this type, here mid-level bank employees and third-year students in Finance. He observed that real claims reduced preference reversal rates to 15 percent from 62 percent in a benchmark hypothetical case.

But while “significant” incentive levels and real-world lotteries with more than two outcomes eliminate preference reversals, these institutions usually do not exist for most environmental goods. This leaves the debate about rational valuation far from over. Many environmental assets, for instance, lack well-defined exchange institutions. As a
consequence, when people are asked to value improvements in the environment, researchers are uncertain as to whether they might continue to reverse preferences because the lotteries are less concrete and no arbitrage exist to discipline bad behavior.

But the distinction between market and nonmarket behavior is not binary. Most people make choices in both thick and non-existent institutions in their daily decisions. The key question therefore is whether the rationality induced from arbitrage in a thick market could spill over to behavior in nonmarket choice. If markets make people reverse preferences less, could this induced rationality spill over to their choices in nonmarket settings?

Cherry et al. (2001) address this question of whether the induced rationality from an arbitraged market could spill over to a second nonmarket choice that would otherwise promote preference reversals. They design a set of experimental treatments to test whether the rationality that markets induce can spill over to nonmarket settings involving isolated individuals, and whether any spillovers that occur are due to preference mutations or to a relaxation of constraints internal to the individual agent.

The experimental design simulates two simultaneous but independent settings—a market and nonmarket setting. The market setting arbitraged preference reversals, whereas the nonmarket setting did not. Treatments were combinations of three variations: (1) real versus hypothetical, (2) arbitrage versus no-arbitrage, and (3) money versus environmental. With real choices, subjects were endowed with money and participated in market transactions that affected their money balances and take-home earnings. Hypothetical choices did not involve money endowments or market transactions. Arbitrage choices identified preference reversals and extracted the potential rent from the irrational behavior. No-arbitrage choices left preference reversals uncontested. Money choices involved lotteries that had winnings and losses in monetary terms. Environmental choices have lotteries comprised of environmental states of the world, e.g., odds of seeing a grizzly bear in Yellowstone, the odds of viewing an eagle or osprey.

With arbitrage, rents from subjects who reversed their preferences were extracted in three steps. The market sold the least preferred/most valued lottery to the subject; next it traded the most preferred lottery for the least preferred lottery; and finally, it bought the most preferred/least valued lottery from the subject. The subject now owned neither lottery, and was left with only a hole in his pocketbook equaling the difference between the stated values for the two lotteries. Arbitrage began in round 6. They also ran a comparative benchmark treatment in which reversals were never arbitraged.

Figure 4 shows the fraction of preference reversals for the marketlike gambles. We see the benchmark no-arbitrage treatment hovering around 25 to 30 percent preference reversals. Once arbitrage is introduced in round 6, we see the rate of preference reversals declines significantly in all three arbitrage-treatments. Rates fall until the reversals were nearly eliminated by round 15. This decline in irrationality is expected given previous work that showed how direct arbitrage eliminates reversals: when irrational choices cost money, people learn to be rational. The question is whether this learned rationality could spill over to the choices in the nonmarket setting.

Figure 5 shows that in fact it does—a rationality spillover exist. Figure 5 depicts behavior in the nonmarket setting, which were choices over lotteries made simultaneously with the market setting. There was no arbitrage in the nonmarket setting.
Nevertheless, after arbitrage was introduced in the other, marketlike setting in round 6, preference reversals also declined in the nonmarket setting. Figure 5 shows that the percentage of reversals in the nonmarket setting declines over the rounds, even though no direct arbitrage took place, the nonmarket choices were hypothetical, and the lotteries became wildlife experiences in Yellowstone National Park. Subjects were apparently taking what they leaned in the market setting, which included arbitrage, and applying it in the nonmarket setting, which did not include arbitrage. People learned that preference reversals can come at cost, and they transferred this lesson to the nonmarket context, even though it did not matter to their pocketbook. Arbitrage is a powerful factor on choices and values.

The next question they asked is whether the people who stop reversing preferences do so by altering their preference ordering, or by changing their stated values for the gambles. The evidence suggests that preference reversals decrease not because the person's preferences changed for risky lotteries, but rather because his stated values decreased. Preferences seemed to hold in tact, people just revised downward their stated value for the riskier lotteries. One could argue that this is good news for valuation work in that preferences are not on some unstable pinwheel fluttering wildly in any new context, but rather people simply overestimated the price the market would pay for a risky lottery. People make these adjustments all the time when they try to buy low and sell high.

This rationality spillover result raises a question about future research in nonmarket choice and rational valuation: how much effort should we allocate between trying to explain biases in behavior versus developing active exchange institutions that induce rationality spillovers which tempers these biases? Currently many economists proceed as if the rational choice paradigm unequivocally holds for nonmarket choices (e.g., the random utility model) despite considerable evidence that many people make biased and inconsistent decisions in these settings. Accordingly, the rationality spillover results suggest that additional effort spent on developing parallel market and nonmarket exchange institutions that generate the rationality spillovers that keep homo economicus intact might be worth the cost. If such lab methods cannot be developed, we might need to develop further a theory of value unaffected by institutional discipline that is more applicable to "irrational" behaviors.

3.3. Synopsis on rational valuation

Standard models of nonmarket valuation presume people make rational choices and reveal consistent values for environmental protection (e.g., household production theory, Hicksian surplus measures). But enough evidence of behavioral anomalies now exists to undercut this presumption of rational choice for people isolated from the repeated give and take with others in an exchange institution (see for example Hey and Orme, 1994). And since exchange institutions do not exist for environmental assets, a person can act as if his value expressions will go uncontested; he is asocial, and need not be accountable to others. Unless one presumes he is a perfect image of an anonymous competitive market that is broad in scope, he may lack the incentives to act in accordance with the utility maximization paradigm and provide the rational valuation one would expect. Without an exchange institution to arbitrage his irrational choices, the unsocialized person can engage in behaviors inconsistent with the paradigm. Unsocialized people fail to
exploit existing gains from trade and engage in behaviors that allow others to exploit these gains (see Akerlof, 1997; Crocker, Shogren, and Turner, 1998).

The lab work on WTP-WTA and the preference reversals illustrates that it is difficult to dismiss as irrelevant the wide range of observed behavioral anomalies. The work also reveals a fundamental difference in conviction about what rational economic valuation means. This difference in conviction underlies the under-debated inconsistency in the valuation literature. At issue is whether one chooses to believe that rational choice is a property of the individual alone or within the social and economic context within which it is embedded. This is a key difference between the debate between the psychology and economic worldview toward rational choice, and the ongoing discussion about which view should dominate how we do research on valuing the environment.

Some researchers with a foot in the psychological camp suggest that we should be interested in “what is in people’s minds, pure and simple.” If one believes that researchers should measure what is in a person’s mind, one could argue that it is acceptable to test economic rationality of people isolated from interactive experience in social exchange institutions. This view might argue that auctions, which are relatively free from any information that would make a person aware of a broader social view, are the most useful demand-revealing tool. For instance, a market-clearing price selected at random does not provide information to the bidder; it is an uninformative number that separates those who pays from those who do not. Advocates of rationality as conscious cognition could then argue that anomalies like the endowment effect matter for valuation.

This viewpoint suggests that much more attention should be devoted to understanding the cognitive processes at work in choice and valuation (see the chapter in the Handbook by Baruch Fischhoff and Shane Frederick). The danger here is that the economists who venture into this cognitive minefield alone will end up fifty years behind the psychologist’s times. In fact, most economic models that attempt to generalize the classic expected utility model do so by adding a needed degree of freedom to explain otherwise unexplainable behavior. The process is to insert an “emotional adder” into the theoretical preference, such as envy, regret, disappointment, malice, and anticipatory feelings, which can then extend the reach of the model (see for example Brennan, 1973; Sugden, 1993; Caplin and Leahy, 2001; Grant et al., 2001). But this gradualist approach to emotional degrees of freedom begs the question as to why add just one emotion—why not several at once if the goal is to more accurately capture the cognitive process at work? One reason might be that few economists probably feel equipped to try and explain restriction on the cross-partial derivative of multiple emotions, say envy and disappointment, or malice and anticipation.

As an alternative, one could choose to work from idea that choices placed in the social context of an active exchange institution are the most informative for economic valuation. This view would argue that tests of economic rationality should not be separated from the interactive experience provided by an exchange institution. Institutions matter because experience can make rational choice transparent to a person—probably the key rediscovery made over three decades of experimental economics research (see for example the theory and examples in Plott, 1996). Asking that choices be made within a social exchange institution, such as second-price and random nth-price auction, in which market prices signal the broader social context is what separates an economics experiment from a psychology experiment. One who supports rationality as a
social construct can argue that a result like the endowment effect is not robust when confronted by the social context of an active exchange institution.\textsuperscript{17}

Both the psychological and economic perspectives can have an important place in understanding how people make choices and state values, depending on the degree of market pressure. One can argue that contact with people making similar decisions helps put in context the economic maxim that choices have consequences. Before one rejects economic rationality as too thin to explain the WTA and WTP gap or preference reversals, one should give the theory the best chance to succeed by testing it within the context that it was motivated—choice within a social exchange institution. The open question is how to make this economic interaction operational for real public goods. The rationality spillover results in Cherry et al. (2001) suggest that a dynamic environment may be a necessary condition. Repeated exposure to competition and discipline was needed to achieve rationality. In becoming rational, people refined their statements of value to better match their preferences. This suggests that efforts to harness the rationality spillover phenomenon could be worth more attention in nonmarket valuation research, especially when trying to value changes in risky events that defines environmental protection (e.g., climate change, biodiversity loss). A design that uses interactive web-based surveys that exploit the power of rationality spillovers could impose more discipline on rational behavior than the typical one-shot questionnaire. How exactly to move the active exchange institutions in the lab into the field in a meaningful way remains an open question worthy of more research.

4. Methods to measure values

We start by examining how the experimental method has been used to directly measure the value of public and private goods, both in the lab and the field. These experiments create unique exchange institutions to elicit values for real goods and services, some with the intention of improving the nature of hypothetical surveys. Lab valuation offers an alternative to hypothetical surveys to elicit values for new products. These experimental markets are actual markets selling goods to people, usually within a stylized setting. The experimental method works to understand, isolate and control how different auctions and market settings affect values, in a setting of replication and repetition. Experiments with repeated market experience provide an environment that allows a person to learn whether sincere revelation of his or her true preferences should be his or her best strategy given the institutional setting.

4.1 Experiments valuing public goods

Since many environmental services are public goods—goods that are non-rival and non-excludable in consumption—it is not surprising that the first use of the experimental methods to elicit values focused on real-world public goods. The early work was in direct response to the claims by some economists that the best approach available was to ask well-structured hypothetical questions about the willingness to pay (WTP) for the public good. Early experiments challenged the claim that these hypothetical valuation statements were good enough approximations even in the face of incentives to distort, or not carefully consider, responses to hypothetical questions.

Bohm’s (1972, 1984) real-world valuation studies of deliverable, non-trivial public goods are the seminal experiments. These two studies addressed whether
incentives to misrepresent WTP produce strong free-rider behavior when WTP for public goods is elicited in the field, and whether combinations of simple, non-truth-revealing mechanisms could attain verifiable approximations useful for actual demand estimation. Consumers knew that the provision of the public good depended on whether their aggregate WTP exceeded production costs, and if the goods were produced, they would pay according to the predetermined payment rules.

In addition, the monitor told participants about the various arguments likely to appear in a public elicitation process. The goal was to mimic a public decision making process in which people had time to discuss the issues with others and would be likely to have heard the media examine the issues and possibly argue in favor of a particular kind of ‘voting’ behavior. Given this objective, attempts were made in the tests to reveal the principal arguments likely to be exposed in such a process and their implications for information about the incentives confronting the respondents.18

In Bohm (1972), a random sample of the inhabitants in Stockholm was asked to participate in a new type of TV rating. Bohm initiated and designed the experiment, which was then run as the Swedish public TV company. Subjects were asked to assign a money value to a new program they could watch if their aggregate WTP were high enough to cover costs. This was in 1969—TV was still exciting in Sweden; the new program was expected to attract wide interest.

Subjects summoned to the premier of a TV program with well-known features and quality were divided into six groups. Subjects in the first five groups were told that if their aggregate stated WTP exceeded a certain cost of showing them the program, they would be given the opportunity to watch the program, prior to which each person would have to pay:

Group I: The WTP stated
Group II: A percentage (as explained) of the WTP stated
Group III: The WTP stated or a percentage (as explained) of the amount stated or 5 kronor or nothing, all four with equal probability
Group IV: 5 kronor, the current average price of a cinema ticket
Group V: Nothing; taxpayers would foot the bill.

Groups I and II, and those in group IV whose WTP fell short of 5 kronor, were exposed to incentives to understate WTP; group V, and those in group IV whose WTP exceeded 5 kronor, were exposed to incentives to overstate WTP. With a dominating number of WTP statements above 5 kronor in group IV (as turned out to be the case), this group can be taken to offer an overstated mean WTP. There were no clear misrepresentation incentives for group III.

Two key results emerged. First, Bohm observed no significant differences between mean WTP in these five groups. This suggests that there were no signs of strong free-rider behavior; the same order of magnitude of true aggregate WTP could have been elicited using any of the five approaches. Second, subjects in the sixth group (VIh, h for hypothetical) were asked to “estimate in money how much you think watching this program would be worth to you” without indicating whether the program would be shown or whether their responses could have an effect on such a decision. Comparing mean WTP responses to this hypothetical question with the pool of the responses to the
five non-hypothetical questions (Bohm, 1994), a Kolmogoroff-Smirnov test showed a difference at the five percent level of significance, i.e., the hypothetical version tended to give an overstatement of the WTP. We return to the question of hypothetical bias in Section 4.1.

In Bohm (1984), the public good was an information package that could only be made available by the government, while the decision to produce the package was to be determined by the actual consumers. In 1982, a Swedish government committee investigating potential operational demand criteria for deciding whether or not to produce costly statistics allowed one particular statistical investment project (providing access to information about certain attributes of housing in Sweden) to be determined by the interval method. If the interval containing the true aggregate or average WTP is small enough, it effectively replaces a true WTP; if large, the output decision is referred back to the politicians.

Potential consumers were identified as the 279 local governments in Sweden. The governments were then randomly split roughly into two groups, A and B. If the good were to be provided, people in group A would pay a percentage of the WTP stated such that the mean payment equaled per capita costs of the project. In contrast, members of group B who stated a WTP of at least SEK 500 would pay a fixed fee of SEK 500 (in 1982, about $100), while those stating a WTP below SEK 500 would be excluded from access to the service. The mean responses from these two groups estimated a lower and an upper bound to the true mean WTP, since Group A had incentive to under-report, whereas Group B had incentive to over-report.

The results indicate that regardless of the two alternative designs of the WTP interval, the natural decision was to have the public good produced. Of the 279 local governments, 274 responded and all of the 130 respondents who qualified for consumption of the good paid the charges. The results reveal a small direct WTP interval—about 7% of the interval mid-point, and a “95% confidence interval” of 38%, 8/9 of which exceeded the cost figure.

The next prominent use of the experimental method to value a public good was Brookshire and Coursey (1987). The nonmarket good in this case was the density of trees in a city park in Fort Collins, Colorado. The objective was to compare and contrast values obtained from hypothetical elicitation methods with values obtained in a market setting. The study consisted of three parts: a hypothetical contingent valuation (CV) study, a field study using the Smith auction (defined shortly), and a lab experiment with the Smith auction.

An artist created a series of pictures depicting different tree densities in the park. A door-to-door CV survey was conducted in which interviewers asked citizens of the community how much they would be willing to pay (or accept in compensation) for an increase (decrease) in the number trees in the park. Each interviewer carried the artist’s renditions of what the park would look like under various different tree densities, either an increase from 200 to 225 or 250 trees or a decrease from 200 to 175 or 150 trees. Citizens did not pay anything.

Brookshire and Coursey (1987) then blended elements of a Smith auction process into their field experiment. The Smith auction has been tested in the lab (Smith, 1980). The results suggest people on average reveal their true induced preferences for public goods—this “on average” is weak if individual incentive compatibility is the goal, but
could arguably be tolerable if efficient provision is the target. The auction collects bids from people. If aggregate bids do not exceed costs to produce the public good, the good is not provided. But if aggregate bids are not less than the cost, each respondent is asked to pay a proportionally scaled-back amount of their bid. People then vote on whether they agree to the price and quantity of the good to be provided. Unanimity is required—one “no” vote kills the process. Brookshire and Coursey did not use the unanimity rule in their field test.

Brookshire and Coursey then examined valuation in the lab with a repeated market-like structure and non-hypothetical transactions. Five trials of the Smith auction were conducted, in which people privately submitted WTP bids or WTA offers. If the subjects’ bids cover the actual cost of providing additional trees, they paid their adjusted bid. The payments were contributed to the Fort Collins recreation department. The divergence between WTA and WTP declined to five-to-one in the lab experiments, considerably smaller than the 75-1 difference in the contingent valuation study. Unfortunately, these WTP-WTA comparisons are easily challenged since the two value measures were inconsistently defined.

4.2 Experiments valuing private risks

Few attempts have been made to use the lab to value directly some private good associated with environment protection. The reason is the difficulty to deliver the private good in the lab, due to the costs. Lab valuation of private goods has primary focused on new food products or processes (see the pioneering work in Menkhaus et al., 1992, and Hoffman et al., 1993). The private-good lab work most closely relates to environmental protection is the experiments to value health risk reductions, as created by foodborne pathogens (see Hayes et al., 1995). While one might ask what foodborne illness has to do with endangered species or climate protection, the lessons learned about how people value reductions in low probability/high severity risks have implications for environmental protection. Like environmental hazards, pathogens pose risks to people. Understanding how people value reductions in risks to life and limb remain a critical part of the valuation question. Lab experiments can address basic questions about how context of exchange affect people’s values for risk reduction in controlled settings. The goal is to understand how actual people react to the consequences of deliverable real risks, and valuing reduced pathogen risks provides a useful case study.

Consider the Hayes et al. (1995) experimental auction market for reduced risk from foodborne pathogens. Hayes et al. construct an experimental auction market to elicit the value of risk reduction. The experimental markets elicit the option price measures of value for five food-borne pathogens—Campylobacter, Salmonella, Staphylococcus aureus, Trichinella spiralis, and Clostridium perfringens. They also ran six treatments to explore how people responded to changes in the risk of illness holding the illness constant. All experiments use money, real food, repeated opportunities to participate in the auction market, and full information on the probability and severity of the food borne pathogen. Subjects ate the food before leaving the lab.

Performed at a meat-testing lab at Iowa State University, the experimental design followed a two stage procedure—first, a pre-auction with candy bars introduced people to Vickrey’s second-price auction. Second, two types of food were introduced, a regular good with the typical odds of being contaminated with a foodborne pathogen, and the
same good stringently screened for pathogens with a low probability (1 in 100 million) of illness. Over twenty trials, the second-price auction elicited bids to upgrade to the stringently controlled food from the regular good. After trial 10, information was revealed about the objective odds and severity.

Figure 6 summarizes the average bidding behavior. Three results emerged. First, subjects underestimate the annual probability of becoming ill from a food-borne pathogen, a result inconsistent with earlier observations on other health risks. Second, the stated values were not robust to changes in the relative risk levels of the five pathogens. Values are fairly constant across pathogens, even given a wide range of risks. Third, examining how people respond to increases in the probability of illness, holding the severity of the illness constant, showed that the marginal option price decreases as risk increases. This observation is consistent with the hypothesis that people will pay more to eliminate the last bit of a risk than they will pay for an equal decrease that still leaves them facing a substantial risk.

The use of the lab to elicit values for reduce risk raises several questions of experimental method. Consider five.

*Does the unique lab environment inflate values?*

The observed premium paid in the Hayes et al. experiments exceeded the expectations of some experts about what people would pay in a real retail market. One explanation might be the novelty of the experimental experience. These auctions are a novel, one-time experience. The concern is that people might be experimenting with their bids by overbidding because the costs of doing so are low.

Theory suggests an alternative explanation for the observed high price premia in the lab—the novelty of the good. Many bidders have never experienced the goods up for auction, e.g., reduction in risks due to some specified or unspecified technology. Here theory says that a bid should reflect two elements of value—the consumption value of the good and the information value of learning how the good fits into his or her preference set (Crocker and Shogren, 1991). Preference learning would exist if people bid large amounts for a good because they wanted to learn about an unfamiliar good they had not previously consumed, because it was unique, or because it was unavailable in local stores.

Shogren et al. (2000) tested these two competing explanations by auctioning off three goods that vary in familiarity—candy bars, mangos, and irradiated pork, in four consecutive experimental auctions over two-weeks. The experimental design followed other private good valuation experiments—second-price auction, repeated trials, posted market-clearing prices. The only difference was that the subjects came back to the lab after several days to let the novelty of the lab experience wear off. The results strongly support the hypothesis that preference learning can explain the high price premia. No statistical change in bids was measured for candy bars and mangos, whereas the price premia for irradiated pork dropped by 50 percent over the four sessions. These findings suggest that people benefit from the information they gain about how an unfamiliar good fits into their preference ordering.
How do posted prices affect bidding behavior?

Lab valuation exercises use multiple trials with posted market prices to provide experience to bidders who walk into these auctions cold. The information sent by a posted market price can help bidders learn about the market mechanism and the upper support of the valuation distribution. But concerns have been raised that market experience “contaminates” bids as posted prices turn independent private values into affiliated private values, especially if people are unfamiliar with the good up for sale (Harrison, Harstad, and Rutström, 1995). List and Shogren (1998) explore this possibility by examining panel data from over forty second-price auctions with repeated trials.

Three results emerged. The market price affects bidding behavior for unfamiliar products, as implied by affiliated private values; the price effect dissipates when bidders receive non-price information about the good or are familiar with the product before entering the lab; and evidence of strategic behavior independent of any price signal still exists. Buyers start bidding low and sellers start offers high, and then bids quickly stabilize after 1 or 2 trials. These results suggest posted prices can influence bidding behavior for unfamiliar products, but the effect dissipates when people have non-price information about the good or are familiar with the good. The results have two implications for lab valuation research: the affiliation of private values can be reduced, if not removed, by providing product information prior to bidding, and a few trials help people learn about the market mechanism.

How does external information affect bidding behavior?

In many environmental controversies (e.g., climate change), the public has had to decide between assertions about levels of risks made by environmental and industry advocacy groups, government officials and scientific experts. The assertions seem at odds to the general public. The experimental method is well-suited to explore how contradictory information affects valuation. One example is the study by Fox et al. (2001) on the demand for reducing health risk from food through irradiation. They explore how favorable and unfavorable information on irradiation affects WTP to control the food-borne pathogen Trichinella in irradiated pork. Using literature currently available to the public, the favorable description emphasizes the safety and benefits of the process; the unfavorable description stresses the potential risks.

Using a random sample of 200 households obtained from a commercial survey company, eighty-seven primary food shoppers were recruited to participate in what was described only as a “consumer economics experiment” in return for a payment of $40.00. Following the experimental design used in earlier work, Fox et al. elicit WTP values in a repeated-trial, second price auction in which the binding trial is chosen at random. They used this feature to examine the adjustment in WTP values that follows the introduction of new information when all participants start from a common informational baseline.

The surprising result is that when Fox et al. presented both positive and negative information at the same time, people were alarmists—the negative information clearly dominated the demand for risk reduction (see Figure 7). This was true even though the source of the negative information was identified as being a consumer advocacy group and the information itself was written in a non-scientific manner. They re-ran this particular experiment four times to ensure that the result was robust, and it was. Negative
reports concerning irradiation had a larger impact on participant preference and values than positive reports—even when the negative reports were unscientific (also see Viscusi, 1997). This asymmetric response to pro and con information can be explained by several theories, including loss aversion, status quo bias, aversion to ambiguity, and Bayesian updating. Which theory best organizes alarmist reactions to new information is an open question.

**How do lab choices compare to retail market behavior?**

Lab experiments are designed to introduce new price and non-price information and then observe the subsequent changes in bidding behavior. But people still know that they are being monitored in a stylized setting, and the range of alternative purchases is more limited than in a retail setting. Shogren, Fox, and Hayes (1999) compare the similarity of lab valuation choices to retail store choices for risk reduction. They also compare choices made in hypothetical surveys. All subjects came from the same small college town and made choices between typical chicken breasts versus chicken breasts irradiated to reduce the risk of food-borne pathogens.

Figure 8 shows that the results from both the survey and experimental market suggested significantly higher levels of acceptability of irradiated chicken than in the retail trial at an equal or discounted price for irradiation. Consumer choices were more similar across market settings at a price premium for irradiation. They observed that in a mail survey and a lab experiment, both of which included information about irradiation, 80 percent of participants preferred irradiated to non-irradiated chicken breasts when they were offered at the same price. When the irradiated product was offered at a higher price, the survey and experimental results predicted market share in a subsequent retail trial remarkably well. About thirty percent of survey respondents, experimental market participants, and shoppers were willing to pay a 10 percent premium for the irradiated chicken, and fifteen to twenty percent were willing to pay a 20 percent premium (also see the in-store experimental method used in Lusk et al., 2001).

While differences in choices across institutions were observed in Shogren, Fox, and Hayes (1999), each of the three decision settings involved unique features and incentives that were absent from the other two settings. Retail purchases involved payment of real money for real products within an environment where thousands of products competed for the consumer’s dollar. Any attempt to collect consumer information in a retail setting was liable to interfere with the realism of the market itself. The goal of the retail setting was to establish the most realistic baseline possible against which one could judge the lab or survey. In contrast, while the experiments also required exchange of real money for real goods, the participants knew they are being monitored and the range of alternative purchases was limited. The survey involved hypothetical decisions, information about irradiation, and people knew that they would not be held accountable for choices they made. Perfectly simulating a retail experience in the lab or a survey so as to control every nuance is unattainable. Rather the goal should be to compare lab and survey choices relative to a real world baseline. The research program that emerges for future work is to explore how comprehensive the lab or survey environments must be to come closer to replicate an actual market outcome.
Does the risk reduction mechanism matter?

People protect themselves from environmental risk through self-protection and self-insurance. They self-protect by curtailing pollution to lower the likelihood that bad states of nature occur; they self-insure by changing production and consumption decisions to reduce the severity of a bad state if it does occur (Ehrlich and Becker, 1972). Risk is endogenous. Self-protection and self-insurance jointly determine risks and the costs to reduce them. And since private citizens have the liberty to adapt on their own accord, a policy maker must consider these adaptive responses when choosing the optimal degree of public mitigation of risk (Shogren and Crocker, 1991). Otherwise, policy actions are more expensive than need be with no additional reduction in risk.

The question is whether risk reduction mechanisms, alone or in combination, affect the value of reduced risk. Shogren and Crocker (1993) designed a set of experimental auctions of sequential substitution between private and collective self-protection or self-insurance. The goal was to reveal implicit preferences for alternative risk reduction mechanisms. The first hypothesis they examined was whether valuation was independent of the sequencing of the private and collective mechanisms. In an early experiment, people preferred private reduction to collective in the single mechanism markets (Shogren, 1990). This preference could remain in the multiple mechanism markets regardless of the sequence of the private or collective auction. The second hypothesis was whether people preferred self-protection to self-insurance. Theory suggests that self-protection that guarantees no risk should be preferred to self-insurance.

The experiment constructed private and collective risk reduction mechanisms by combining two auctions for self-protection and self-insurance. The private auction is a Vickrey sealed bid, second price auction where the highest bidder secures the 100% risk reduction and pays the highest losing bid. The Vickrey auction has been promoted as a possible elicitation device for contingent valuation because of its well-known preference-revealing properties. The collective auction was a modified Smith (1980) sealed bid auction, in which the cost to reduce risk equaled the sum of the subjects’ expected consumer surplus. Costs were not public information. If the sum of collective bids exceed the costs of provision, the mean collective bid was posted as the reigning price. Unanimity was required such that any one subject could veto collective risk reduction, which in retrospect probably reduced efficiency (see Banks et al., 1988). Collective reduction was also rejected if the summed bids were less than costs.

Shogren and Crocker’s results suggest the combined risk reduction mechanism matters to valuation, but not as much as when the mechanisms are separated. When private reduction was available prior to collective action, people place a greater value of the private action. But if private action was accessed after collective action, values were independent of the action. They also rejected the hypothesis that self-protection was valued more highly than self-insurance. Di Mauro and Maffioletti (1996) confirmed this result, finding no evidence that the mechanisms constitute a "frame."

In contrast, Shogren (1990) found that the mechanism used to reduce risk mattered; reducing risk by altering the probability or severity of an undesired event through a private or a collective mechanism generated significantly different values. Generally, the upper bound on value was the private reduction of the probability of an undesired event; the lower bound was the collective reduction of severity. The
preference for a mechanism occurred for several reasons including the sense of personal control, the marginal productivity, and whether the odds or severity are altered.

Finally, Kask et al. (2001) found evidence in a contingent valuation survey on valuing reduced risks from dioxin that people had a clear preference for the reduction mechanisms. Nearly everyone avoided the collective insurance approach. About 40 percent of the subjects preferred the collective protection approach to all others. But in the mortality sample, private approaches were slightly more preferred to collective approaches. Overall, the results say that people do prefer both the level of risk reduction and how the risk is reduced, but the results are sensitive the economic context.

4.3. Synopsis on direct valuation

As a complement to traditional tools like econometrics, hedonics, and contingent valuation, experimental methods directly applied to questions of valuation in the field and the lab continue to develop into useful tools for eliciting consumer values for real decisions in a controlled environment. They entail real payments and binding budget constraints: experiments use auctions to sell goods for money, albeit within a stylized setting. More importantly, experimental designs can isolate and control the market setting to address specific questions. After decades of work, the experimental procedures have passed at least one critical test: they have enabled researchers to learn things about behavior that would have been impossible to discover from any of the alternative procedures. For example, when faced with both positive and negative information about new food technologies, consumers react as if they had received only negative information.

The contextual design of laboratory auctions also matters. First among these is having a subject make actual decisions such that they exchange money for real goods. Real exchanges of money and goods are aimed at inducing sincere behavior and punishing irrational behavior. Lab research has also found seemingly trivial aspects of the protocol to be important. Paying participants of an experimental auction prior to the auction rather than afterwards reinforces the monetary incentive. Further, reminding the participants that bidding zero is acceptable mitigates any assumption that positive bidding is expected. Auction multiple goods at once also can be done, but designing demand-revealing auctions becomes more complicated.

We also know that repeated trials in laboratory auctions can affect bidding behavior. The repetition allows bidders to bid again after observing the market-clearing price for a new product that has just entered into their opportunity set, a nontrivial event for a distinctive commodity like irradiated meat. Some bidders react to the posted market price by changing their bids in the next round; others do not. Repetition has both plus and minus for value elicitation. If the goal is to elicit each bidder’s value independent of signals from the market, repetition is not recommended (see Harrison, 1996). Evidence suggests that posted prices can influence bidders who are inexperienced or unfamiliar with the good in question. A one-shot auction provides these raw values “uncontaminated” by market information. But if the goal is to understand how a person’s valuation of a good reacts to market experience and new information—provided by either a market price or a written description, repeated trials provide the flexibility for a researcher interested in informed bidding behavior. Repeated second-price auctions, for example, eliminated the divergence between WTP and WTA in the case of market goods with close substitutes. The impact of information can also be estimated by observing
bidding behavior before and after the release of information. Similarly, the impact of marketing and taste tests can be estimated with repeated auctions.

We have also learned that there are limitations to what can be achieved with experiments. Collecting refined information about the value of risk reductions can be difficult—people have general preferences for clusters of goods rather than specific goods. But if this is the case in a sterile laboratory it is likely to hold for survey valuation research too. Subtle changes in the experimental procedure such as whether participants were paid ahead of time, whether the reported WTP or WTA, or whether they knew the market-clearing price can significantly impact the results. Bids for new, unfamiliar goods can be unrealistically high when participants viewed them as a novelty. And over time as designs are refined, improved reality-based valuation experiments seem likely to develop into a useful method for applied economists.

5. Methods to testbed surveys

The idea to use public opinion surveys to assign values the natural environment took flight with Davis's (1963) groundwork on hypothetical iterative bidding games. The path-breaking idea seems simple enough now—just ask people how much they would be willing to pay for a hypothetical change in environmental protection. A researcher could construct a hypothetical “contingent” market containing features that parallel real-world markets and institutions. This pseudo-market could be manipulated to conform to the problem at hand. Quantity and quality dimensions such as temporal context, spatial dimensions, property right entitlements, and uncertainty could be varied to reflect specific policy questions. Ideally, a well-structured contingent valuation survey would let each person solve his own trade-off problems and reveal his \textit{ex ante} valuation for the hypothetical change. And as discussed in the Handbook chapter by Michael Hanemann, researchers around the world now use surveys to elicit stated preferences for all categories of environmental issues.

But it is no secret that using surveys to elicit values has drawbacks and detractors.\footnote{Using the lab to overcome these drawbacks was why some environmental economists were drawn to the experimental method in the first place. In the 1980s, Cummings et al. (1986) promoted the lab as just the tool to strengthen surveys before they are implemented in the field. Their idea was to use the lab as a testbed for hypothetical surveys. The lab could be used to help researchers understand how people learn about incentive systems before these designs were used in the field. Lab testbeds have more control and can be repeated. The lessons learned from isolating and controlling potential biases prior to field implementation could improve the accuracy of the final survey. Ideally, as Coursey and Schulze (1986, p. 48) note, survey practitioners would "walk away from the laboratory with a 'best set' of questionnaires," accurately revealing preferences for the environmental asset.}

We now consider two broad areas in which the experimental method is being used to explore behavior, with the ultimate goal being to improve hypothetical survey work—hypothetical bias and framing effects.
5.1 Hypothetical bias I: Does it exist?

The classic jab at valuation surveys remains: “ask a hypothetical question, get a hypothetical answer” (A. Scott, as quoted in Bishop and Heberlein, 1986). Despite numerous attempts to dismiss it, survey work has never shaken this troublesome gibe from four decades ago. McCloskey (1985, p. 181) notes that "one can get an audience of economists to laugh out loud" by suggesting to send out a questionnaire. Economists prefer observations from actual market behavior to public opinion surveys, which in some eyes, has tarnished contingent valuation results. They hold up results like the Forsythe et al. (1992) experiment on the Iowa Political Stock Market. The IPSM allows traders to buy and sell portfolios of shares in political candidates in a double auction market. Their results show that the IPSM outperformed opinion polls on the 1988 presidential election, despite the judgment bias of individual traders.

Plus, the trial and error approach that dominated earlier valuation research left some observers concerned; they worry that no rational economic theory exists to help explain hypothetical choices based on intended behavior. The lack of replication to verify survey results has also troubled some observers. We now consider studies that have explored the question of hypothetical bias. We begin with Bohm's (1972) classic experimental lab study, which compared bids in hypothetical and actual experimental markets that elicited subjects’ stated value to preview a Swedish television show. His results suggest people moderately overstate their actual values when asked a hypothetical question.

The tendency to overstate actual cash commitments was supported in the seminal field experiments by Bishop and Heberlein (1979, 1986) on the value of goose and deer hunting permit experiments in Wisconsin. In the goose experiment, they found that average and median actual cash values to hunters were significantly less than those suggested in the hypothetical question. The deer hunting experiment generated the same result—overvaluation in the survey relative to actual cash outlays.

Subsequent research has generally supported the observation that hypothetical values exceed actual WTP (e.g., Seip and Strand, 1992; Neill et al., 1994; Frykblom, 1997; Balistreri et al. 1998; Spencer et al. 1998). Exceptions exist—some experiments have found no significant difference in real and hypothetical behavior (e.g., Dickie et al., 1987; Loomis et al., 1996; Sinden, 1988; Smith and Mansfield, 1998). And a few other studies found mixed results. Battalio et al. (1990), for instance, tested for differences in responses for hypothetical verses real payoffs. They found systematic and significant quantitative differences over real and hypothetical choice questions. Subjects were more risk averse with real payoffs than with hypothetical payoffs. But even though there were significant quantitative differences between hypothetical and real payoffs, the qualitative findings with respect to real verses hypothetical payoffs were similar.

The mass of the evidence, however, suggests that the average person exaggerates his or her actual willingness to pay across a broad spectrum of goods with vastly different experimental parameters (see Harrison and Rutström’s review, 1999). For instance, the ratio of hypothetical-to-actual overbidding, which ranged from 2.2 to 3.5 for baseball cards, falls between the ratios observed for irradiated pork, watercolor paintings, maps, and other goods that ranged from, 1.0 to 10.0 (see Diamond and Hausman, 1994). They reinforce the argument that people tend to overstate their actual WTP when confronted with hypothetical questions.
Researchers have spent less energy on understanding the relationship between the real and hypothetical WTA measure of value. The lab evidence from this relatively small lot of studies is mixed. Bishop et al. found that Wisconsin goose hunters overstated their actual WTA to sell goose-licenses; deer hunters in a sealed-bid auction understated their actual WTA to sell deer-permits, while hunters in a dichotomous choice institution overstated their real WTA. Coursey et al. (1987) found that people overstated their actual WTA to experience a drop of the sucrose octa-acetate (SOA); Smith and Mansfield’s (1998) field survey suggests that real and hypothetical WTA statements for the opportunity to spend time in a second set of interviews on an undisclosed topic are statistically equivalent. Again these results suggest that the real-hypothetical WTA gap might be case-specific, conditional on the good and the context. List and Shogren (2001) calibrate real and hypothetical WTA estimates elicited for consumer goods in a multi-unit, random nth-price auction. Their results suggest that people understated their real WTA in the hypothetical regimes, framed both as demand and non-demand revealing exchanges.

5.2 Hypothetical bias II: Calibration

Another relevant question suited for the lab is whether values elicited from hypothetical surveys can be calibrated to reflect the amounts individuals would pay for proposed programs. The National Oceanic and Atmospheric Administration’s (NOAA) blue-ribbon panel recommended that hypothetical bids from surveys be calibrated using a “divide by 2” rule, unless these bids can be adjusted using actual market data (NOAA, 1994). The NOAA rule has served as an ad hoc placeholder to motivate more research into the nature of calibrating hypothetical and actual values.26

To illustrate, consider List and Shogren’s (1998) field experiment that compares bidding behavior in a hypothetical and actual second-price auction for baseball cards—a good with many characteristics favorable for a calibration exercise including familiarity, the ability to deliver, and an intangible quality. Three samples were run—one card, 1 card among 10, and 1 card bid on by sportscard dealers presumed more experienced with the market than the general population. While the results support the view that people overstate actual bids, the calibration function estimated to correct for this exaggeration is both good- and context-specific, i.e., other goods and market experience affect the calibration function.

Calibration research is continuing to develop in the lab. One alternative method is the CVM-X method (Fox et al., 1998). CVM-X works in four steps. Step 1: use a survey to elicit hypothetical values for the good in question. Step 2: bring subsamples of the survey respondents into the laboratory and elicit real bids for the actual good in an incentive-compatible auction that employs real goods, real money, and repeated market experience. Step 3: estimate a calibration function relating the auction market bids of the subsample to the hypothetical survey bids. Step 4: use the estimated calibration function to adjust the values of the survey respondents who did not participate in the laboratory auction. CVM-X could be a cost-effective tool that combines the advantages of the stated preference, contingent valuation method (CVM) and experimental auction markets (X). The method could be used to increase the validity and accuracy of surveys while broadening the scope of nonmarket valuation in the lab.
The CVM-X application studied by Fox et al. (1998) is the reduction in health risk from the parasite *Trichinella* achieved with food irradiation. Irradiated foods are not yet widely available in the U.S., and most people are unfamiliar with the process—which gives it a feature common to many nonmarket environmental goods like biodiversity. Nearly two hundred randomly selected households participated in the survey. They were asked the maximum they would be willing to pay to upgrade from their less-preferred sandwich to their sandwich of choice in an open-ended elicitation question. At the end of the interview, participants who were pork eaters were asked if they would be interested in participating in a consumer economics experiment.

In the lab experiment, participants were assigned to one of two treatments—the irradiated or the non-irradiated treatment. The experimental auction procedures followed those in Hayes et al. (1995)—stage I was the candy bar auction (second-price); stage II was the food auction. Figures 9 and 10 show individual bids between the survey and trial 2 in both treatments. The results suggest that an upward bias in hypothetical bids exists, and that the lab can be used to correct for this bias, but the calibration function might be commodity-specific.

Researchers are usually more interested in public goods and programs, items that lack the deliverability of reduced health risk. Future research should explore whether a private good is a reasonable proxy for a public good, and whether a systematic method exists to cluster goods into classes of calibration functions. Consider two other attempts to calibrate hypothetical and real values across goods. The cross-commodity “bias function” approach of Blackburn et al. (1994) rests on the presumption that bias for a good in one context is measurable and transferable to another good in another context. Blackburn et al. use discrete choice data on subjects who participated first in a hypothetical and then in a real valuation setting for one private good. A multinomial logit model is used to explain the pattern of responses (yes-yes, yes-no, no-no) and to determine whether the bias in hypothetical responses is systematically related to socioeconomic characteristics. The evidence to support transferability of the bias function to a second private good is inconclusive given the relatively large standard errors of the estimated coefficients (also see Swallow’s discussion their calibration method, 1994).

Harrison et al. (1997) calibrate open-ended values for wetlands protection using two bias functions: one to account for the downward bias due to free riding and the other to account for hypothetical bias. The free riding bias function is measured in a comparison of two real valuation situations for a nature calendar, one of which features public provision of the calendar. This bias function is used to calibrate both real and hypothetical bids for the preservation of local wetlands with the corrected bids used to estimate the hypothetical bias function. Both calibration functions, free riding followed by hypothetical, are then applied to bids elicited for national wetlands preservation. While intuitively appealing, the approach rests on little tested assumptions about transferability of bias functions between different contexts: real versus hypothetical values, private versus public goods. The approach also makes an implicit and untested presumption that these biases are additive.27

Cummings and Taylor (1999) propose a different path to remove the hypothetical bias—a so-called *cheap talk* survey design. Their idea is that one might be able reduce hypothetical bias before it starts by the choice of wording in the survey. They propose
that simply telling a respondent, before he answers the valuation question, about the hypothetical bias issue might remove the bias. If people are told that they usually exaggerate their reported values, the hypothesis is that this frank revelation will cause him to reflect and revise his hypothetical bid downward toward the value he might pay. Cummings and Taylor run sixteen different treatments that compare stated values with and without cheap talk to a benchmark treatment that elicited actual contributions to four different public goods (e.g., contributions to the Nature Conservancy in Georgia). Their results suggest that the cheap talk design worked to make responses to hypothetical questions indistinguishable to responses involving actual cash payments, and that this effect was robust across changes in script and public good, with one exception.

While promising, neither the calibration nor the cheap talk research provide a substantive explanation as to why people react the way they do when making hypothetical choices. An open question is why respondents reduce hypothetical values when told that people have a tendency to inflate hypothetical values. Perhaps there is some deep cognitive reason or maybe it is just the Hawthorne effect at work—subjects want to meet the expectations of the experimenter. It is unanswered questions like this that keep fears over hypothetical behavior alive in nonmarket valuation debates. One cannot argue with Mansfield’s (1998, p. 680) point that “the power of the calibration model could be improved by a better understanding of how individuals answer valuation questions, including the traits or attitudes that inspire individuals to give more or less accurate answers.” No camp has thus far provided a convincing axiomatic explanation as to what creates or removes the wedge between intentions and actions. The lack of an analytical framework increases the odds that this discussion will stagnate into “did not, did too.” The debate will continue until a robust and parsimonious theory emerges as to why the wedge occurs and whether it can be predicted and controlled systematically.

5.3 Framing Effects I: Surrogate Bidding (Scope/Embedding)

Most researchers involved in valuation surveys know that how a question is asked is as important as what is asked. These so-called framing effects show up in many forms in valuation questions: the description of the good, the payment mechanism, the description of available outside options and substitutes, and reminders of budget constraints. Whatever the reason, the main concern is that a hypothetical survey on the environment does not measure what it is supposed to measure, or it measures values left uncontrollable by the experimental design. Consider two areas of experimental work that can be used to better understand how framing affects values—surrogate bidding (this section) and incentive compatibility (the next section).

In the state of the art review of valuation surveys, Kahneman (1986) pointed out that hypothetical questions might produce surrogate bidding, also called embedding or insensitivity to scope. Surrogate bidding exists when hypothetical (or non-hypothetical) values for a specific good reflect preferences for environmental phenomena in general; or when values are insensitive to changes in the quantity or quality of the good. A person asked to state his WTP to protect one endangered Wyoming Toad might be revealing his value to protect all toads, or to protect every endangered species. For example, McClelland et al. (1992) found that up to one-half of the reported values for a specific environmental charge can be attributed to surrogate values. The basic instinct behind surrogate bidding is reflected in the language used to describe this behavior: warm glow,
the purchase of moral satisfaction, insensitivity to scope (quantity or quality), part-whole bias, and embedding. These terms suggest that people might be insufficiently sensitive to the continuum of goods or services they are valuing; a person’s stated value captures his general feeling toward this class of goods. Surrogate bidding means that values for specific goods cannot be relied on as accurate indicators of preference because it is unclear what values are being elicited.

Most tests of surrogate bidding have been performed using data from contingent valuation surveys (see for example Hammitt and Graham, 1999). A few surrogate bidding experiments examining non-hypothetical choices have been run in the lab to gain more control through real economic commitments. Kealy et al. (1990) is the first classroom experiment. They designed an experiment to test the reliability of survey methods given the general concern about surrogate bidding, focusing on the effect of familiarity with the good. The experiment used students from an undergraduate psychology class at Colgate University. The students were asked their WTP to contribute to the New York Department of Environment Conservation for decreasing acid rain in the Adirondacks. The objective was to compare behavior for a private good with which the students had experience to behavior for a public good with which they had little experience. Some subjects were warned that they would be asked to actually make the contributions that they claimed they were willing to pay; others were not warned.

Of the warned students, nearly 95 percent paid an average WTP of about $18. Of those not warned, 27 percent refused to pay, and those that did pay had an average WTP of about $5. This result suggests that people were valuing something other than what the study purported to measure. Many people who found a legitimate reason to back out of the contribution did so, which suggests the questionnaire might have been eliciting surrogate preferences toward the Adirondacks, not just contributions for a specific (but vague) program to decrease acid rain.

Three ensuing lab experiments also point toward surrogate bidding. Boyce et al. (1992) designed a lab experiment to elicit non-use values for environmental protection. Their experimental design asked people to value a houseplant (a Norfolk pine) that would be destroyed unless they protected it by buying the plant. People paid more when they were explicitly told that the plant would be killed if they did not buy it. But it is unclear however what value this design was capturing—the non-use value for the houseplant or a statement of rejection to a situation in which someone would take a morally objectionable action.

Hayes et al. (1995) also explore whether surrogate bidding existed in the value of risk reduction. They compared the bidding behavior from each foodborne pathogen treatment to the bids from a treatment that combines the risks of all the pathogens—a 1 in about 46,000 chance of illness per meal from at least one of the five pathogens. Their hypothesis was that similar bidding behavior across treatments would not contradict the idea of surrogate bidding. The results suggest that surrogate bidding existed for reduced health risk: bids for a cluster of pathogens were indistinguishable from bids for specific pathogens. Using a contingent valuation survey, Hammitt and Graham (1999) reproduced the Hayes et al. study, and found the same insensitivity to probability.

Bateman et al. (1997) found similar results for restaurant meals. They considered the existence of surrogate bidding, or part-whole bias, in the lab. They used an incentive
compatible mechanism to auction off vouchers for parts of a restaurant meal. They elicited values for the parts and the whole were elicited. They observed that the sum of the parts exceeded the whole, again supporting the idea that values seem to reflect general preferences for a good.

Some observers, however, remain unconvinced that surrogate bidding, or insensitivity to scope, is a major problem for valuation. Citing evidence from the hypothetical survey literature, Carson et al. (2001, p. 183) say that “[p]oorly executed survey design and administration procedures appear to be a primary cause of problems in studies not exhibiting sensitivity to scope.” If the poor design tag is also aimed at lab results, the proposition can be tested. Researchers can replicate these experiments under new economics conditions to test the robustness of the results, which at the present time suggests some level of surrogate bidding. This is the beauty of the experimental method—if you do not agree with the findings or you doubt certain design features, you are free to run your own experiments to refute or replicate earlier work.

A new lab experiments could also be construct to test whether arbitrage could remove the tendency for surrogate bidding. A clever design could be created in which the monitor buys and sells different bundles of these goods such that the person is left with less for more. He then might have the needed economic incentive to think through his stated valuations for changes in both quantity and quality.

5.4 Framing II: Incentive Compatibility

Now consider framing and incentive compatibility. Can one defining an exchange institution as incentive compatible, or demand-revealing, even though the exchange is hypothetical? Experimental economics has long explored the nature of incentive design in the provision of public goods like environmental protection. These experiments explore how behavior differs from the received theory which predicts that rational, self-interested people free-ride off the contributions of others, and thus markets will fail to provide the optimal level of the public good making alternative organizations necessary. While several variants exist on the public good experiment, the basic design captures the idea that the efficient outcome is to cooperate but the dominant strategy is to free-ride.30

For valuation surveys, however, it is unclear whether people will (1) tell the truth, (2) free ride if they think they might pay for the good, or (3) overstate their bids if this increases the chance of getting the good at a low cost. While some survey supporters have interpreted the lab results as saying that free riding is not a problem (e.g., Mitchell and Carson, 1987), the evidence is mixed: some people free ride, while others coordinate actions (see Ledyard 1995). The findings suggest that incentive compatible mechanisms can be used to increase efficiency, but truth telling is not the predominant observed behavior (see Taylor, 1998).

The early public good experiments prompted valuation researchers to compare these mechanisms to the standard survey method of asking a direct WTP question. Bennett (1987), for instance, examined how a modified Smith auction eliminated strategic behavior in the provision of a hypothetical public good relative to a direct question approach (recall the Smith auction was also used in Brookshire and Coursey, 1987). He found that direct statements of hypothetical values lead to over-bidding of induced value, whereas the Smith auction lead to under-bidding. He suggested that the
most promising avenue to estimate values in real-world cases is to combine the direct statements of value with the lab experiments to tease out truthful bidding.

Prince et al. (1992) used the lab to consider the properties of another public good provision mechanism—a contribution game mechanism. They wanted to know whether this mechanism could be used to increase the accuracy of surveys. The contribution game pulls together a group of people and asks them to contribute to a public good. Subjects know the group size, cost of the good, and the payoff to each individual. Individuals choose how much to contribute. If contributions fall below costs, the good is withheld and contributions returned. If contributions exceed costs, the good is provided, with surplus funds going to the supplier. The Prince et al. results suggest that surveys that use the contribution game mechanism can be incentive compatible and might overcome several observed cognitive problems, at least for an identifiable subset of the population.

Rondeau et al. (1999) explored the properties of a provision point mechanism for public good provision. Their design used induced values and an environment designed to mimic field conditions. The one-shot provision point mechanism works as follows: A relatively large group (50+) is pulled together, and each participant is endowed with an initial balance of, say $6. Each person then enters a bid indicating how much of the $6 he or she will contribute to a group investment fund. The sum of the bids contributed to the investment fund must equal or exceed an “investment cost” for the investment to be made on behalf of the group. This investment cost is predetermined but unknown to the bidders. If the sum of group bids equals or exceeds the investment cost, each member of the group receives a randomly assigned private payoff that may differ from other group members’ payoffs. Each person’s earnings depend on his or her bid and on whether the investment cost for the group was reached.

Two possible outcomes were considered: (1) if the investment cost was not met by the group, the full amount of a person’s bid was refunded; or (2) if the investment cost was exactly met or exceeded, each person received his or her payoff from the group investment. And if the bids exceeded the cost, all of the excess was rebated to the group. The rebate was directly proportional to the amount of one’s bid relative to the total amount of the group’s bids. Thus, if someone’s bid was 30 percent of the sum of bids for the group, his rebate would be 30 percent of the excess bids.

The results suggest that the provision point mechanism is demand revealing “in aggregate” for a large group with heterogeneous preferences. Rondeau et al. argue that these results suggests a relatively simple mechanism could be used in the field to elicit preferences leading to the efficient provision of a public good.

But here is the rub with this provision point mechanism. The mechanism is argued to be incentive compatible at the group level. Unfortunately, the economic theory of incentive compatibility is defined at the individual level—each person should have incentive to bid sincerely. Incentive compatibility is an individual concept, not a group concept. If the group reaches the provision point because low value bidders overstate WTP and high value bidders understate WTP just means that the mechanism might be lead to an efficient outcome on average; it does not mean it is incentive compatible. The idea of using this mechanism in field surveys begs the skeptic’s question of whether a mechanism that is ‘just right’ on average is close enough for valuation work.
Another incentive compatible issue is the claim that the dichotomous choice (DC) method used in survey work is demand revealing. The DC method asks a person to say yes or no to paying a fixed dollar amount for a hypothetical change in some good or service. If the DC method is incentive compatible in a survey, it implies that a person would answer the same way when faced with a real cash commitment. Cummings et al. (1995) use the lab to test this proposition. They construct treatments to compare behavior between- and within-subjects. The between-subject treatments draw two samples from the same population. Each person in the first group was asked whether he or she would pay actual cash for the good at a given price; people in the other group were asked the same question, but it was hypothetical and no binding exchanges were made. The within-subject treatments asked a person if hypothetically he or she would buy the good at a given price, and then the good was provided and the question was repeated, except now the commitment was real. The result that hypothetical responses overstated the actual value paid by the subjects caused Cummings et al. to reject the thesis that the hypothetical DC method was incentive compatible.

Additional experimental work has followed up on the demand revealing nature of the DC question. For example, Brown et al. (1996) explored valuation for an environmental good under four conditions: real cash payments to direct and DC questions, and hypothetical payments to direct and DC requests. Based on the results, they concluded that the DC mechanism seems to be more of an issue than hypothetical behavior. The DC question causes people to say yes too often, irrespective of whether the choice is hypothetical or real. Frykblom and Shogren (2000) sold people a volume of the Swedish National Atlas called Environment using either a DC choice or within Vickrey auction. All sales were real, nothing hypothetical. Their results suggest that while the DC format induced greater mean WTP than the Vickrey auction (e.g., 81.41SEK vs. 71.02SEK), the difference was not statistically significant. Following the quasi-experimental method used in the Bishop-Heberlein goose studies, Champ et al. (1997) compared how a sample of Wisconsin residents react to hypothetical referenda versus actual DC contributions to a road-removal program at the North Rim of Grand Canyon National Park. Their results again show that the hypothetical donations were poor predictors of actual contributions.

Johannesson et al. (1998) considered the idea that sometimes a yes does not mean yes. They compared a DC question with real cash commitment, but this time only an "absolutely sure yes" is counted as a yes. Their results suggest while a standard hypothetical "yes" overestimates the real "yes", an absolutely sure hypothetical yes did not (also see the similar results in Blumenschein et al., 1998). The question mark with this line of research is that while one might wonder what a yes means, no theory exists to suggest a systematic reason why one yes means something different than another yes. Such "what is 'is'" debates on the semantics of what a choice really meant seems has to require more attention to the details in the cognitive psychology literature on language and imagery, a new area for most economists.

Along the lines of the DC experiments, Cummings et al. (1997) examined whether a hypothetical referenda over the provision of a public good is incentive compatible. The valuation literature has advanced the thesis that a reference should be demand revealing because if people prefer the public good at the hypothetical posted price, they should vote "yea"; otherwise, they vote “nay.” There should be no reason to
misrepresent their preferences, or inferred willingness to pay. Cummings et al.’s experiment was an attempt to empirically test the thesis of whether behavior under real and hypothetical referenda differed or not.

The experiment worked as follows. Thirteen distinct groups were pull together, and told the following: if at least half of the group voted for the public good for a $10 donation per person, the good would be provided and everyone in the group would make the donation; otherwise, no donations would be made. The treatment across was whether the vote was for real or hypothetical. The real-voting groups were told the money would be collected immediately after the vote if the “yea’s” have it; the hypothetical-voting groups were told that no money would be exchange, but they should make their decisions as if their vote and donation were real. The public good was $2N (N - people in a voting group) bilingual “citizens guide” booklets that described areas in Albuquerque that overlie contaminated groundwater, how to test for contamination, and what self-protection actions are available if a well is contaminated.

Based on their analysis of the data, Cummings et al. rejected their thesis that voting behavior was independent of the use of a real or hypothetical referendum. They then conclude with a reprimand for anyone who would presume incentive compatibility for hypothetical referenda without testing empirically for it first. In response, Haab et al. (1999) take up the empirical gauntlet by reevaluating the Cummings et al. data. They relax the presumption of homoskedasticity across treatments because real votes (with real opportunity costs) are expected to have less variability than hypothetical votes. They show that once the error scales are correctly identified, they rejected Cumming et al. original null hypothesis—no significant difference was observed between real and hypothetical referenda, and that both voting rules are quite noisy signals of WTP. Haab et al.’s constructive re-evaluation clearly tosses the question of (in)sincere voting in hypothetical referenda back to the lab for more examination.

5.5. Synopsis on testbedding

This brief review illustrates the movement made over the last two decades toward achieving Coursey and Schulze’s (1986) aim of using the lab to testbed field surveys. More researchers, but not many, now use the lab to better understand how people react to incentives provided by valuation questions prior to field application. The one sign that the lab valuation work has impacted valuation research is that the approach now has its own formal critics. They stress that a laboratory testbed has its own problems that restrict the applicability of findings for survey work. And as usual for experiments, the debate is usually over the lack or loss of experimental control. If concerns over uncontrolled incentives in the lab make sense, lab results can be challenged as openly as the hypothetical surveys the lab research was designed to address.

Bishop and Heberlein (1986), for example, considered the lab work on the Vickrey auction as a “red herring.” They pointed out that while a second-price auction might be demand-revealing in theory for a private good, most environmental problems are public goods. Moreover, complicated incentive-compatible auctions for private or public goods might simply increase costs, increase confusion, and decrease response rates for a survey.

Another criticism of lab work involves uncontrolled outside options, and this affects real and hypothetical bidding. Suppose a researcher thinks that the lab is
producing untainted values for a real private good, whereas bidding behavior is conditioned by the market price of the same or similar goods sold outside the lab. The existence of such unmeasured prices for outside options has been argued to explain differences in laboratory behavior when hypothetical payments exceed real payments for the same good (Harrison, 1992; Smith, 1994). Hypothetical bids, so the argument goes are not too high. Rather, real bids are too low because they are truncated at an actual market price set outside the lab. If this is the case, insights about valuation that arise from lab results might be less instructive than previously believed.

Smith and Mansfield (1998) advance this dissent to form the most cogent set of criticisms to date. They first challenge the negative concerns raised about contingent valuation work because they were “based on diverse experimental evidence drawn from small, specialized samples involving purchase of private goods and contributions for goods with public attributes.” They go on to rebuke the experimental literature for providing “little systematic treatment of how the circumstances of choice influence the analyst’s ability to describe preference relations from choice functions.” They conclude by noting that no objective benchmark exists to determine a value’s “‘realness’” or the “degree of ‘hypotheticality’” elicited in any experiment that does not explicitly control private preferences.

These are reasonable concerns that deserve answers. Consider each point in turn. First, the diversity of experiments and sample populations should not be seen as a weakness of the lab, but rather as a strength. The weight of the evidence from both student and lay subject pools in numerous settings involving many different goods, elicitation methods, and contexts run by many different analysts point in one direction—hypothetical values usually exceed actual values. It is difficult to dismiss the experimental evidence as “hazy” in light of Harrison and Rutström’s (1999) review that shows 34 of 39 tests revealing hypothetical bias, ranging from 2 to 2600 percent.

Second, while more research into behavior will always be desirable, it is amiss to argue that little systematic treatment exists on how context affects choice and values in the lab. The literature reveals the patterns that have emerged for valuation: frames matter (e.g., Kahneman et al., 1990), the nature of the good matters (e.g., Hayes et al., 1995), information matters (e.g., Cummings and Taylor, 1999; Fox et al., 2001), exchange institutions matter (e.g., Bohm, 1972; Rutström, 1998), market experience matters (Shogren et al., 1994), information on the market-clearing price can matter (Fox et al., 1998), and substitutes and complements can matter (List and Shogren, 1998). Rather than dismissing experimental valuation research as a hodgepodge of unrelated and ad hoc treatments, it is more constructive to note that one experiment begets another experiments. The lab valuation literature continues to follow the classic experimental strategy: start simple, and add complexity slowly to understand what factors matter, and why. As peer-reviewed evidence continues to accumulate, a clearer and more definitive picture will emerge about the context of choice and valuation.

Finally, Smith and Mansfield are correct in that no one will ever know with absolute certainty whether someone is bidding sincerely in the lab given private and wild preferences. Two strategies exist: surrender the search and follow Bohm’s (1984) idea of the interval method, in which a “value interval” is created by splitting the sample into those with an incentive to free ride and those with an incentive to overstate values; or keep on the trail by testing and retesting the reliability of incentive compatible
mechanisms given preferences controlled or induced by the monitor (see Forsythe and Isaac, 1982). For instance, the second-price auction has been a popular mechanism because it is demand-revealing in theory, relatively simple to explain, and has an endogenous market-clearing price. But the auction has its problems, even with induced values. People bid insincerely, especially bidders who are off the margin (bidders whose value is far below or above the market-clearing price). The auction seems not to engage low-value bidders who think they will never win, which suggests the auction is unreliable if one is trying to measure the entire demand curve for a real-world good.

But rather than debating about criteria for “realness,” such problems point the way toward potential solutions. For instance, Shogren et al. (2001b) examine the potential for the random $n$th-price auction to attempt to engage these otherwise disengaged off-the-margin bidders. The auction now has a random and endogenously determined market-clearing price. Like the Becker-DeGroot-Marschak (BDM), randomness is used to engage all bidders because everyone has a chance to buy a unit of the good, and the endogenous price guarantees that the market-clearing price retains some relation to bidders’ private values. The lab evidence suggests that the auction can induce sincere bidding behavior in theory and practice. Does this mean the success of the auction translates automatically for private preferences? No, but it does suggest that one can use the lab to help increase the confidence in the tools we use. While advocates of the lab, including myself, may have overstated the case, it is clear that using the experimental method can help improve field valuation through repetition and the accumulation of evidence both for and against alternative auctions and contexts. What has emerged thus far is that the relationship between real and hypothetical values seems to be both good- and context-specific, and that attempts to bridge the gap statistically might be difficult since the gap appears to be non-transferable across commodities.

6. Concluding Remarks

Thirty-five years ago, Allen Kneese (1966, p. 87) wrote: “optimal rules, standards, or other techniques for controlling environmental quality must result from analysis of values, contrary to the usual approach which is still narrowly focused on physical effects and objectives. Research in the economic values associated with environmental management has made significant progress along some lines, but has barely begun to shed light on many difficult problems.” Today the valuation literature has nearly four decades of enlightenment under its belt. And each step forward has uncovered even deeper questions about the behavioral underpinnings of valuation. Within each new insight dwells another doubt, a progression that plays to the strength of the experimental method and mindset. The beauty of the experimental method is that questions of increasing difficulty and complexity can be addressed step by step in the lab or the field. By combining observations with first principles the ‘sooty empiric’ using the experimental method in valuation work has shown that economic context does matter to valuation. A review of the experimental method applied to valuation helps pinpoint what seems to matter the most: the nature of the good, information sets, exchange institutions, market experience, price information, and outside options. The experimental method applied to valuation can highlight what matters and why by creating a controlled setting in which treatments are replicated under similar conditions. Systematic work of replication and validation will help to increase the acceptance of valuation research. The
lab or the field can be used to test how well general theories hold in specific cases, or how additional complications affect both hypothetical and real stated values.

Debates will continue over control and context, realism and simplicity, role-playing and anonymous economic agents, and instinctual reactions and experienced behavior within and along side markets. Walking the fence line between realism and control generates lively exchanges because both are concepts that can be achieved within broad limits defined by human subjects—realism to one researcher is a lack of control to another; control to one represents an artificial edifice to another. Regardless the goal should be to create a valuation environment in which people make motivated choices with economic consequences.

And even if the appropriate mix of control and realism can be agreed on, even motivated respondents who are answering valuation questions can have preferences, beliefs, and skill that differ from those of the traditional *homo economicus* (see for example Mazzotta and Opaluch, 1995). Some exchange institutions are robust in an efficiency measure to these behavioral differences (e.g., the double-oral auction); others are not. The challenge is to understand what makes an institution robust, how much reality can be added to the lab without losing perceptible control, and whether the person’s choices are motivated by economic considerations or not—tasks especially relevant to valuation research which for the most part presumes rational decision makers.

The pitch for the experimental mindset in valuation is two fold: the method imposes a discipline on researchers to test and explore the key behavioral elements that underlie a valuation problem; and once identified, the method gives researchers a pliable tool to sample and resample institutional rules to help find systematic behavioral patterns in both motivated and unmotivated statements of value. The experimental mindset for environmental valuation holds regardless on whether one chooses to believe that rational choice is a property of the individual alone or is embedded within the social context of the decision. The choice, however, sets on a distinctive path. If one believes that researchers should measure the value that exists in a person's mind at the time the question was asked, one could make the case that valuation should be isolated from interactive experience in real social exchange institutions. Here isolated value statements of value reveal preferences free from any information that would make a person aware of a broader social view towards the good on the auction block. Less economic context and weaker exchange institutions to punish irrational behavior creates an environment in which the wide range of psychological contextual cues can take over choices and values. When psychological context is all that matters, economists do not have a comparative advantage in understanding why and how; it becomes the psychologist’s job to understand the range of contextual stimuli and responses.

In contrast, if one believes that choices and values emerge in the social context of an active exchange institution (e.g., Arrow, 1987; Becker, 1962), estimates of value should not be separated from the interactive experience provided by an exchange institution, either directly or indirectly. Institutions matter because experience can make rational choice transparent to a person (see Smith, 1991). Asking that choices be made within a real social exchange institution is what separates an economics valuation experiment from a psychology valuation experiment. The institutional context matters for economic choices and values. Environmental economists can advance this research direction further by addressing the reality that most people make allocation decisions in
many institutional settings each day—markets, missing markets, and nonmarkets. The question to take to the lab then is: how do the rational choices and preferences formed within active exchange institutions affect nonmarket behavior and stated values for environmental protection? This inquiry matters because it is the contact with others who are making similar decisions that puts in context the economic maxim that choices have consequences, and values have meaning. Valuing environmental protection should not be the exception.
Footnotes


2 Early experiments looking at individual choice, utility and preference emerged in the published literature in the 1940s and 1950s. The work of Mosteller and Nogee (1951) was an early attempt to measure the nature of preferences with the experimental method. They were interested in using a laboratory experiment to measure “the value to individuals of additional money income” (p. 371).

3 The experimental method as applied to economics continues to have its open-minded critics. A good example is Ariel Rubinstein (2001), who accepts the idea that experiments can help theorists protect themselves from misguided intuition. Rubinstein, however, points out that “we rely on our own integrity to such extent that economics (so I am told) is one of the only major academic professions that lacks a code of ethics….It is hard to avoid the thought that the lack of clear standards in experimental economics and the burdens placed on experimental research procedures actually serve as barriers to entry into the field” (p. 625, 627).

4 The triad is a simplified version of that found in Mount and Reiter (1974) and Smith (1982).

5 The other four conditions include nonsatiation, dominance (net benefits of participation are positive), privacy, and parallelism (behavior in one choice transfers to another similar choice, holding conditions constant). If you relate this to the involuntary unemployment and credit rationing literatures, adverse selection says that one should not use a subject willing to work for such low wages because he is most likely going to shirk anyway. But defining the elements that must exist in an experimental design to satisfy each of these conditions is arguably a judgment call. Appeals to “common sense” and decrees of “sloppy designs” rest in many cases on the ability to measure whether subjects are making motivated choices or not, which can differ across people and context, in often subtle ways. A good example is the Kachelmeier and Shehatta (1992) experiment that examined risk preferences over high stakes lotteries (three times their monthly wages) in China. They conclude that monetary incentives seem to induce a measurable response only under extremely high rewards; otherwise the level of reward seemed to make little difference.

And while spirited, these debates have yet to rise to the rank reported in the February 27th 1918 edition of the Cloquet Pine Knot (as quoted in Fahlstrom, 1997, pp. 85-86): “A couple of Finns who had evidently been pouring out (or rather pouring in) copious libations in honor of St. Patrick, got into a theological discussion in Osborne’s saloon on Saturday afternoon last. One of them attempted to emphasize his remarks with a 10 pound cuspidor, the true inwardness of which, running down the face of his adversary, added nothing to his comeliness. Whereupon five of his countrymen joined in a fight and for a time the battle raged with great vigor….Then Policeman Olin appeared, and selecting the cuspidor conversationalist…he marched him off to the calaboose, where he passed the night. Sunday morning he awoke sober and penitent, and considering that he had been sufficiently punished, he was released with an admonition to be less emphatic in his discussions in the future.”

6 Some have argued that hypothetical surveys are experiments because they satisfy the sufficient conditions of a controlled microeconomic experiment (Bergstrom and Stoll,
The term *choice experiment* has been used to describe work that uses hypothetical pair-wise choices to estimate marginal values for a good.

Read the exchanges between Cummings et al. (1986) and the outside reviewers (e.g., Ken Arrow, Daniel Kahneman, Sherwin Rosen, Rick Freeman, Rich Bishop, Kerry Smith) during the state of the art assessment of the contingent valuation method. Design features Cummings et al. presented as a tool for control, others argued were simplistic and artificial (see Bishop and Heberlein, 1986).

Economists who run experiments today can identify with linear programmers and econometricians from the past. Fifty years ago, an economist might need a day to run a regression or linear program given the limited capacity and computing power, especially if the run was done by hand in an auditorium packed with people using adding machines and passing slips of paper. Once the runs were complete, one either lived with the results or went through the process of collecting more funds for more time-consuming runs.

Experimenters should be wary of the classic Hawthorne effect: the presence of monitors affects subjects’ behavior, which then confounds results and makes interpretations untrustworthy. The effect is named for the episode at the Hawthorne plant of the Western Electric Company in Chicago when monitors were asked to track how job performance affected productivity. With the monitors watching over their workers’ shoulder, they improved their job performance and productivity (Benson, 1994). By explicitly looking for more productivity the monitors induced the productivity they were looking for. This concern over the self-fulfilling prophecy has also been called the "pygmalion effect," named after the sculptor Pygmalion of Cyprus who fell in love with his own creation—an ivory statue of the ideal woman named Galatea. The goddess Venus eventually granted Pygmalion’s prayers to bring Galatea to life, and the couple lived happily ever after. See, however, Jones (1992) who found little evidence of the Hawthorne effect in the original study.

See the surveys by Cummings et al. (1986), and Horowitz and McConnell (1999).


See for example Camerer’s (1995) overview on procedural invariance and preference reversals.

See for example Slovic and Lichtenstein (1983) and Tversky et al. (1990).

See for example Karni and Safra (1987) and Fishburn (1988).

Cherry and Shogren (2001) explored whether the rationality impact of arbitrage extended to a set of diverse decision-making tasks over preferences for gambles. Their results suggest that arbitrage in one setting can crossover to affect choices over unrelated tasks: stated values for safer food dropped by 20 to 50 percent, and the frequency of the Allais paradox falls by half. But as expected, the frequency of the more distinct Ellsberg paradox remained the same. They also found that the type of arbitrage, real market-like experience or a simple oral description, did not affect the results.

Harrison (1992, p. 1441) makes the point with more aplomb: “the analysis of financially unmotivated behavior can be left to those most equipped intellectually to handle it: psychologists….We arguably have no useful business fussing around in an attempt to make sense out of unmotivated behavior.” But Harrison’s message can be pushed too far for some aspects of environmental protection. Most environmental goods exist outside the market but along side others goods bought and sold in a market. It is an empirical question whether the emotive preferences for the environment affect or are
affected by the rational preferences for wealth and durables. Environmental economists need to confront the mix of transaction modes that people actually operating in when asked to value environmental protection. It is unclear exactly how “motivated” or “unmotivated” people really are within this mix. It is an empirical question that must be tested in environmental that have markets, missing markets, and no markets (recall Figure 2).

Some observers argue a person’s interaction with a market will affect the preferences the researcher is trying to measure. The notion is that markets do more than just allocate resources, they also affect the “evolution of values, tastes, and personalities” (Bowles, 1998, p. 75). This viewpoint suggests that if the idea is to use social interaction through markets to keep consistent preferences reigned in, the researcher might instead be changing the nature of these preferences by exposure to the market.

Although it is difficult to speculate on the nature of an equilibrium state of such information, it was presumed that certain incentives to misrepresent WTP would be widely known and talked about, and that the organizers of the ‘referendum’ would try to counter these incentives by referring to the ‘duties’ of citizens participating in this kind of public process or the meaninglessness of conducting ‘referenda’ of this type if voters gave in to such incentives. Bohm’s dialogue with his subjects can be interpreted as an early attempt at what is now called “cheap talk,” which is discussed in Section 5.2.

Note that Banks et al. (1988), however, showed that unanimity actually decreases the efficiency of the Smith auction because the rule cause people to lose money.

During the same time period, Bennett (1983) emulated Bohm’s work, using a variant of the Smith auction both for a real public good, the communal viewing of a film.

See the Handbook chapter by W. Kip Viscusi and Ted Gayner on risk assessment and management.

Affiliation exists when one bidder who values the good highly increases the chance that other bidders will also put a high value on the good.

Shogren (1990) also observed that with repeated market trials, the initial bid was a significant predictor of the final experienced bid; the implication is that an initial bid, adjusted for learning, could reflect the value of reduced risk in an experienced market.

See Cummings et al. (1986) and Carson et al. (2001) for discussions of the various potential biases and the likely impact on the validity and reliability of contingent valuation.

The concern with hypothetical survey questions can be traced at least as far back to Allen Wallis and Milton Friedman’s (1942) critical review of Thurstone’s (1931) seminal experimental measurement of an indifference curve. They argue that “[f]or a satisfactory experiment it is essential that the subject give actual reactions to actual stimuli....The response are valueless because the subject cannot know how he would react” (p. 180).

Others offer an alternative vision. Randall (1996, p. 200), for example, states that: "[t]he calibration issue, it seems to me, is an audacious attempt to promote a Kuhnian paradigm shift....I would argue vigorously that the essential premise is unproven and the question is therefore premature and presumptuous. The proposed new calibration paradigm is at this moment merely a rambunctious challenger to the dominant external validation paradigm."

Another classic contingent valuation bias arises with some types of calibration procedures—strategic bias. If respondents know that their hypothetical values will be
calibrated with actual values later, they might have incentive to act strategically and inflate their initial values.

28 Early on, Bishop and Herberlein’s (1986) called for more attention to the social psychology literature. But rather than digging into an unfamiliar discipline, some observers have a more pragmatic view toward cheap talk. If cheap talk works to lower values, does it really matter why it works? While nice to know, we do not need to know the theory of physics to ride a bike.

29 See the exchange between Kahneman and Knetsch (1992) and Smith (1992).

30 Suppose four people are each endowed with $5. Each person is asked to make either a private contribution of either $0 or $5 to a collective program. Every $1 contribution returns $2 to the group—$0.50 to the contributor and $1.50 to the other three subjects. The efficient outcome is for everyone to contribute $5, but the dominant strategy is to contribute $0 since the private net return is negative.

31 It is unclear why one would want to create an environment in which the subjects were uncertain about the investment costs if the mechanism is intended for real-world collective decisions. A significant fraction of the people in the real world would probably want to know the likely costs of the project up front before they commit resources to this public sector plan.

Acknowledgements

Thanks to the USDA/ERS, NSF, and USEPA for helping support this work. Thanks to Bjorn Carlén, Todd Cherry, Tom Crocker, Ron Cummings, Sean Fox, Shane Frederick, Peter Frykblom, Dermot Hayes, Terry Hurley, Jayson Lusk, Greg Parkhurst, Charlie Plott, Laura Taylor, and Jeff Vincent for their helpful comments. Thanks to Peter Bohm for his comments, encouragement, and support. All views remain my own.
References


Conant, J. (1951), Science and Common Sense (Yale University Press, New Haven, CT).


List of Figure Legends

Figure 1. Experimental Economic Triad (Mount and Reiter, 1974; Smith, 1982)

Figure 2. An Intersection of Experimental Methods

Figure 3. Ratio of Mean WTA & WTP by Auction (Mugs)

Figure 4. Preference Reversal Rates in the Market Setting

Figure 5. Preference Reversal Rates in the Nonmarket Setting

Figure 6. Average Bid to Exchange Risky Sandwich for Riskless Sandwich

Figure 7. Average Bid to Exchange Meat for Irradiated Meat before and after Information

Figure 8. “Yes” to Stated Price for Irradiated Chicken (percent)

Figure 9. Comparison of Hypothetical and Auction Bids for Irradiated Pork

Figure 10. Comparison of Hypothetical and Auction Bids for Non-Irradiated Pork
Table 1—Summary of Experimental Design Parameters

<table>
<thead>
<tr>
<th>Design Parameter</th>
<th>Original Experimental Designs</th>
<th>Shogren et al. (2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens, pens, &amp; mugs</td>
<td>Candy bar, sandwich, &amp; mugs</td>
<td>Candy bar &amp; mugs</td>
</tr>
<tr>
<td><strong>Initial Monetary Endowment</strong></td>
<td>None</td>
<td>$3: candy bar $15: sandwich or mug</td>
</tr>
<tr>
<td><strong>Number of Trials</strong></td>
<td>Varied between 3-7</td>
<td>5: candy bar</td>
</tr>
<tr>
<td></td>
<td>20: sandwich</td>
<td>10: mugs</td>
</tr>
<tr>
<td><strong>Retail Price Information</strong></td>
<td>Provided for some treatments</td>
<td>None provided</td>
</tr>
<tr>
<td><strong>Subject Participation</strong></td>
<td>In-class</td>
<td>Voluntary</td>
</tr>
<tr>
<td><strong>Number of Subjects per Session</strong></td>
<td>Varied between 30 &amp; 44 Simon Fraser U.</td>
<td>12 to 15</td>
</tr>
<tr>
<td></td>
<td>Iowa State U.</td>
<td></td>
</tr>
<tr>
<td><strong>Auction Institution</strong></td>
<td>Becker-DeGroot-Marschak Mechanism (BDM)</td>
<td>Second-price Auction (SPA)</td>
</tr>
</tbody>
</table>
Figure 1. Experimental Economic Triad (Mount and Reiter, 1974; Smith, 1982)
Mainstream Experimental Economics [Markets]

Non-Economics Motivated Experiments [No Markets]—Psychology, Social theory, Organization theory

Experimental Environmental Economics [Missing markets]: Risk, Conflict, Cooperation, Control, Valuation, and Prosperity

Figure 2. An Intersection of Experimental Methods
Figure 3. Ratio of Mean WTA & WTP by Auction (Mugs)
Figure 4. – Preference Reversal Rates in the Market Setting
Figure 5. – Preference Reversal Rates in the Nonmarket Setting
Figure 6. Average Bid to Exchange a Risky Sandwich for Riskless Sandwich

Campylobacter, Salmonella, Staphylococcus aureus, Trichinella spiralis, Clostridium perfringens, All pathogens

$\text{Pre-Information bids} \quad \text{Post-Information bids}$
Figure 7. Average Bid to exchange Meat for Irradiated Meat before and after Information

- Pre-Information bids
- Post-Information bids

Bars represent the average bids for different types of information.
Figure 8. "Yes" to Stated Price for Irradiated Chicken (Percent)
Figure 9. Comparison of Hypothetical and Auction Bids for Irradiated Pork

Bids for Irradiated Pork (N=58)

Trial 2 Bid  Hypothetical Bid
Figure 10. Comparison of Hypothetical and Auction Bids for Non-Irradiated Pork

Bids for Non-Irradiated Pork (N=19)

- $\text{Trial 2 Bid}$
- $\text{Hypothetical Bid}$

Participant

0 0.5 1 1.5 2 2.5

$\text{Trial 2 Bid}$ $\text{Hypothetical Bid}$