

# Experimental Evidence on Alternative Environmental Valuation Methods

by

Glenn W. Harrison<sup>†</sup>

June 2005

Forthcoming, *Environmental and Resource Economics*

## Abstract

Experimental methods are central to assessments of environmental valuation approaches that are operationally meaningful. Existing lab experiments focus attention sharply on the neglect of hypothetical bias. They also offer constructive solutions to correct this bias, and beg for validation in field experiments.

<sup>†</sup> Professor of Economics, Department of Economics, College of Business Administration, University of Central Florida (E-mail: gharrison@bus.ucf.edu). I am grateful to Richard Carson, Maribeth Coller, Ron Cummings, Ron Harstad, Bengt Kriström, J. Clay Lesley, John List, Tanga McDaniel, Steven Nape, Elisabet Rutström, Melonie Sullivan, the editors, and two referees for discussions and comments. All jokes and errors are mine.

Experimental methods have taken a dominant position over the academic debate of the validity of methods of environmental valuation. There are good reasons for this development. The most popular methods of environmental damage assessment involve a hypothetical survey. In order to determine the extent of any hypothetical bias in the values elicited one must know the true values of the respondent. This is the domain of controlled, theory-driven experiments. Although there are many other uses for experimental methods, the assessment of the extent of hypothetical bias is, without doubt, the most important area of application in the field of environmental valuation.<sup>1</sup>

I critically review recent developments in the use of laboratory experiments to compare stated and revealed preferences. The term “stated preferences” will be taken as referring to responses that do not entail any real economic commitment by the subject *or* real economic consequences (e.g., delivery of the commodity being valued). “Revealed preferences,” in contrast, entail a real economic commitment, a real consequence, or both.<sup>2</sup> In each case the response could be ordinal or cardinal, and could refer to two or more alternatives.

To focus the issues, I will restrict analysis to studies that consider revealed preferences in a controlled laboratory setting where one has some *a priori* basis from theory to believe that the responses being elicited reflect true valuations. This simply reflects the desperate need in the literature to have some empirical point of reference. There is an unfortunate tendency to view different sets of responses abstractly as simply “different data generation processes,” and to engage in “black box” statistical modeling of any differences. Even worse, the modeling often occurs in the error term, virtually ensuring a capitulation of any coherent theoretical perspective on differences. The motto proposed here is *festine lente*: hasten slowly. If we explain and mitigate the simplest differences, then we have no business adding noise by throwing in lots of different sources of un-motivated data.

The goal is a critical review of experimental procedures and suggestions for further use of experiments in environmental valuation. There have been many sloppy experiments in this area, and invalid inferences have been casually drawn from them. These erroneous conclusions come from critics

---

<sup>1</sup> Experimental economics has also made contributions in other areas, but space constrains coverage. In particular, the analysis of “part-whole” valuation issues, free riding, discount rates, and risk aversion are all active areas of research by experimenters that impact environmental valuation.

<sup>2</sup> These extremes may obscure the possibility of intermediate cases, where the subject does not make a real economic commitment but where there are potentially real consequences of his actions. Such might be the case, for example, in a field survey in which the subject believes that his responses to a payment question will entail no payment from him directly but might influence somebody else’s payment to him (e.g., the *Exxon Valdez* settlement amount).

and supporters of the use of lab methods alike. The goal is to *weigh* the recent evidence, as a guide to sorting out the contributions that are likely to be of lasting methodological value. Cummings and Harrison [1994] offer a critical review of the older experimental literature, and Shogren [2004] describes his experience with experiments and valuation.

The most important finding from recent experimental work is that “hypothetical bias” is an important and robust factor in valuation tasks, and is not reliably mitigated by varying basic elicitation procedures. Section 1 reviews these findings in detail. However, there is striking evidence that lab experiments might provide a constructive basis for mitigating or calibrating these biases. The experiments that offer these positive results involve the careful combination of stated and revealed preferences, either through deliberate experimental design or through statistical procedures. Section 3 reviews the first of these approaches, which I call “instrument calibration.” Section 4 reviews the latter approach, which I call “statistical calibration.”

The first conclusion is that the problem of hypothetical bias cannot be circumvented by simply changing the valuation mechanism. When early research showed that open-ended valuation methods exhibited hypothetical bias, a common response was to say that such bias was expected and that it was only dichotomous choice settings that would be free of such bias. When subsequent research showed that the dichotomous choice setting was not free of hypothetical bias, the response was that the binary context should be that of a referendum in order for bias to disappear. Subsequent research showed that the binary referendum setting was not free of hypothetical bias. The response was to say that the valuation setting should be “stated choice,” which is just an application of revealed preference logic to capture bounds on indifference curves. While the experimental evidence is not complete with respect to stated choice as the “savior” of field valuation methods from hypothetical bias, it does not take a rocket scientist to see the defensive trend in the brief history of this literature.

The second conclusion is somewhat more constructive. It is a series of recommendations as to “best practice” in the use of experimental methods to elicit homegrown values. From all of the criticisms, important and trivial, there emerges a series of sensible and defensible proscriptions as to how one should conduct an experiment that does not suffer from those criticisms.

A general theme, running throughout, is to identify a low-cost research program which uses lab experiments as a basis for operationally meaningful communication between critics and supporters of

different methods for valuing the environment. Although a major thrust of previous experimental work has been directed at undermining the false confidence that hypothetical survey proponents assert in their method, the next stage in this research is likely to emphasize the complementary nature of field and lab valuation exercises. Concrete proposals for how this complementarity may be realized are offered.

## **1. Hypothetical Bias**

The class of experiments of concern here deal with the elicitation of “homegrown values” rather than the experimenter-imposition of “induced values.” In most respects the two types of experiments are conducted in the same manner. Subjects are typically recruited from college populations in a convenience sample, the experimenter provides instructions in the procedures, a trial run of procedures is often conducted using some good other than the one under study, and the subjects are not deceived. In general the commodities being valued have been either ordinary private goods, such as juicers or sandwiches, or contrived public goods, such as contributions to a fund that purchases some “public good.”

One departure from traditional lab experiments has been a concern with subject heterogeneity, which is mitigated by collection of information on observable individual characteristics and their use in statistical analysis. This procedure has become fashionable in lab experiments recently, but was standard in experiments eliciting homegrown values since one expects preferences to be subjective. Of course, in the case of homegrown values preferences may be subjective because some subjects have better knowledge of the particular good being valued, and this is also a matter for control using survey questions about purchase experience or attitudes.

Another departure is that the instructions often tell subjects the “demand revealing” properties of the institution being used. This is because one does not always want to test the joint hypothesis that the subjects infer those properties logically and that they have such and such a value; oftentimes one is only interested in the latter inference, so knowledge of the incentives for truthful revelation would be a confound.

A final departure has been concern with the possibilities of arbitrage between the lab and the field. Since the commodities being valued in the lab are often readily available in the field, this may cause some “affiliation” of values or “censoring” of lab responses, as discussed below. In a conceptual sense, this is

perhaps the most important difference between induced value experiments and homegrown value experiments (Harrison, Harstad and Rutström [2004]).

### 1.1 Evidence on Hypothetical Bias

The experimental literature on hypothetical bias, which refers to differences in response between settings in which the consequences are hypothetical or real, may be easily summarized. The designs are transparent, and the inferences clear. It is a testimony to the power of a “conventional wisdom” in the environmental valuation literature that so much has needed to be written on so little. Harrison and Rutström [2004] offer a more systematic, quantitative review.

#### *A. Open-Ended Elicitation*

Neill et al. [1994] opened the recent debate on the existence of hypothetical bias by conducting experiments with Vickrey auctions for private, deliverable commodities. Their auctions were “one shot,” and the instructions contained language explaining in simple terms the dominant strategy property of the institution. Each subject was asked to bid on a small oil painting by a justifiably unknown Navajo artist.

The most interesting design feature of these experiments is the attempt to differentiate a Generic Contingent Valuation Method (CVM) valuation task from a Hypothetical Vickrey Auction. The former amounted to a relatively unstructured request for the subject to just state the maximum amount of money they would be willing to pay for the painting. No allocation or provision rules were discussed, although the instructions made it clear that the question was hypothetical. The latter was identical to the Real Vickrey Auction treatment except that the instructions were minimally changed to reflect the hypothetical nature of the transaction. The goal of this design was to see how much of the hypothetical bias might be due to the hypothetical nature of the economic commitment, and how much might be due to the absence of structured instructions using a potentially demand-revealing institution.<sup>3</sup>

---

<sup>3</sup> Loomis, Brown, Lucero and Peterson [1996] illustrate the dangers of not using scripts that reflect demand-revealing institutions. They have one hypothetical session with no details of the sealed-bid auction, one hypothetical session with an interesting treatment (asking subjects not to just state what they thought the fair market value was) and details that indicate a first-price sealed-bid auction, and one real session with details that indicate a first-price sealed-bid auction. The first-price sealed-bid auction is not demand revealing: optimal bids in a symmetric Nash Equilibrium depend on risk attitudes, knowledge of the number of other bidders, expectations about the bidding strategies of all other players, and a myriad of other off-equilibrium factors. For example, with risk-neutral bidders in a Nash Equilibrium, optimal bids are below true values. The fact that their second session had bids lower than the first treatment is therefore confounded by their use of the first-price procedures and

Their results were clear: the culprit was the lack of a real economic commitment in either of the two hypothetical institutions. Real valuations were about 40% of the hypothetical transactions. The fact that the Hypothetical Vickrey Auction included a great deal of “cheap talk” about auction procedures, and the implied incentive to report truthfully, apparently had no effect on behavior.<sup>4</sup>

Rutström [1998] examines more closely the question of experimental procedures in open-ended elicitation. Does an English auction give different results than a Vickrey auction, even if both are supposed to be demand-revealing? It does, in ways that are explored in detail in Harrison, Harstad and Rutström [2004]. Does it matter if subjects are paid different amounts to turn up? Yes, but in ways that can be easily corrected for by just collecting information on observable socio-economic characteristics and allowing for those in cross-sample comparisons. Does it matter if subjects receive random monetary endowments at the beginning of each session? Yes, but not as much as one might think, and in ways that can again be corrected for easily using appropriate covariates in the data analysis.

There are three experimental design issues raised by popular practices employed by many experimenters, such as Hayes et al. [1995], Hoffman et al. [1993], Shogren, et al. [1994] and Bohm, Lindén and Sonnegård [1997].

Design Issue #1: Affiliated Values Many experimenters use repeated Vickrey auctions in which subjects learn about prevailing prices. This results in a loss of control, since we are dealing with the elicitation of homegrown values rather than experimenter-induced private values. To the extent that homegrown values are *affiliated* across subjects, we can expect an effect on elicited values from using repeated Vickrey auctions rather than a one-shot Vickrey auction.<sup>5</sup> There are two reasons why homegrown values might be affiliated in such experiments.

The first is that the good being auctioned might have some uncertain attributes, and fellow bidders might have more or less information about those attributes. Depending on how one perceives the

---

their treatment.

<sup>4</sup> The concept of cheap talk is discussed in more detail in Section 2. It refers to any instruction which has no real consequences for the payoffs to given responses, such as “Please tell the truth when you respond to this question.” Of course, cheap talk can have strategic consequences, such as in coordination games.

<sup>5</sup> The theoretical and experimental literature makes this point clearly by comparing real-time English auctions with sealed-bid Vickrey auctions: see Milgrom and Weber [1982] and Kagel, Harstad and Levin [1987]. The same logic that applies for the one-shot English auction applies for a repeated Vickrey auction, even if the specific bidding opponents were randomly drawn from the population in each round. The concept of affiliated values refers to values that are correlated. The concept of homegrown values refers to values that subjects have when they come to an experiment, and are distinct from any values induced on them by the experimenter. Harrison, Harstad and Rutström [2004] review these concepts further.

knowledge of other bidders, observation of their bidding behavior<sup>6</sup> can affect a given bidder's estimate of the true subjective value to the extent that they change the bidder's estimate of the lottery of attributes being auctioned off.<sup>7</sup> What is being affected here by this knowledge is the subject's best estimate of the subjective value of the good. The auction is still eliciting a truthful revelation of this subjective value, it is just that the subjective value itself can change with information on the bidding behavior of others.

The second reason that bids might be affiliated is that the good might have some extra-experimental market price. Assuming transactions costs of entering the "outside" market to be zero for a moment, information gleaned from the bidding behavior of others can help the bidder infer what that market price might be. To the extent that it is less than the subjective value of the good, this information might result in the bidder deliberately bidding low in the experiment.<sup>8</sup> The reason is that the expected utility to bidding below the true value is clearly positive: if lower bidding results in somebody else winning the object at a price below the true value, then the bidder can (costlessly) enter the outside market anyway. If lower bidding results in the bidder winning the object, then consumer surplus is greater than if the object had been bought in the outside market. This argument suggests that subjects might have an incentive to strategically misrepresent their true subjective value.<sup>9</sup>

The upshot of these concerns is that unless one assumes that homegrown values for the good are certain and not affiliated across bidders, or can provide evidence that they are not affiliated in specific settings,<sup>10</sup> one should avoid the use of institutions that can have uncontrolled influences on estimates of true subjective value and/or the incentive to truthfully reveal that value. Specifically, repeated Vickrey

---

<sup>6</sup> The term "bidding behavior" is used to allow for information about bids as well as non-bids. In the repeated Vickrey auction it is the former that is provided (for winners in previous periods). In the one-shot English auction it is the latter (for those who have not yet caved in at the prevailing price). Although the inferential steps in using these two types of information differ, they are each informative in the same sense. Hence my remarks about the dangers of using repeated Vickrey auctions apply equally to the use of English auctions.

<sup>7</sup> See Harrison, Harstad and Rutström [2004] for an explanation of these pathways.

<sup>8</sup> Harrison [1992] makes this point in relation to some previous experimental studies attempting to elicit homegrown values for goods with readily accessible outside markets. Coller and Williams [1999], Harrison, Harstad and Rutström [2004] and Harrison, Lau, Rutström and Sullivan [2005] illustrate its application using statistical procedures for censored responses.

<sup>9</sup> It is also possible that information about likely outside market prices could also affect the individual's estimate of true subjective value. Informal personal experience, albeit over a panel data set, is that higher-priced gifts seem to elicit warmer glows from spouses and spousal-equivalents.

<sup>10</sup> List and Shogren [1999] purport to provide such evidence. Unfortunately, their analysis only looks at median prices, rather than the entire distribution of prices or mean prices. One would not expect affiliation or censoring to affect median prices as much as mean prices, since it would be expected *a priori* to have more of an effect on the "tails."

auctions are not recommended as a general matter.<sup>11</sup>

Design Issue #2: Dominant Strategy Adoption The second problem with many popular experimental designs is that subjects are not told the demand-revealing properties of the auction. There is considerable evidence from experiments with Vickrey auctions with induced values that subjects need some time and practice to learn the dominant-strategy bidding property in these auctions.

When the objective of the experiment is to *test for* the propensity of subjects to use this strategy it is obviously not particularly wise to tell subjects about the strategy. But when the objective is to *elicit* homegrown values there seems little point in adding the noise of requiring subjects to learn about this strategy in real-time. For this reason, most experiments using Vickrey auctions to elicit homegrown values have employed instructions explicitly<sup>12</sup> telling subjects about the dominant strategy.<sup>13</sup>

An alternative approach is to use a series of training experiments in induced value settings to give subjects a chance to learn about the dominant strategy property.<sup>14</sup> Kagel, Harstad and Levin [1987] and Harstad [2000] suggest that the best procedure here would be to train subjects in an English auction setting rather than a Vickrey auction setting. The reason is that subjects appear to learn the dominant strategy property more quickly<sup>15</sup> in the former: as price goes above their true subjective value, each tick of the clock forces every surviving subject to consider the logic of bidding in excess of true value, whereas in a Vickrey auction this logic only disciplines behavior when the subject wins the auction. Of course, these effects are clearly identified in all standard theoretical models of bidding behavior in these institutions, but they are persuasive and consistent with observed experimental behavior.

Design Issue #3: Asymmetric Budget Constraints The third potential problem with many experimental designs relates to the question of budget constraints, and whether or not subjects perceive the budget constraint to be binding. Horowitz [1991; p.320] and Neill et al. [1994] discovered that some subjects in their experiments did not behave as if the real economic commitment asked of them was actually credible. That is, some subjects behaved as if the cash payment requirement was hypothetical even

---

<sup>11</sup> If there existed some way of untangling the true subjective value from the observed bid, then these problems would not be severe. However, the current state of theory does not encourage one to rely on any particular model of these influences.

<sup>12</sup> Boyce et al. [1989] apparently employed similar instructions, but did so orally and so there is no way to know whether the language they used accurately conveyed the strategy.

<sup>13</sup> See Neill et al. [1994] and Rutström [1998].

<sup>14</sup> See the painting experiments of Neill et al. [1994], in which an induced value trainer of several Vickrey auctions was employed along with (previously prepared) oral instructions as to the dominant strategy logic in that trainer.

<sup>15</sup> And in a way that appears to transfer across institutions, as illustrated by Harstad [2000].

though they had been *instructed* that it would be real. Their response was to modify their Vickrey auction procedures so as to require subjects to put cash (or cash-equivalent) in an envelope prior to their bid being considered feasible, and for this to be checked by a monitor for consistency with the bid amount. The envelopes of all losers would be returned unopened, and the winner would receive back the difference between his bid price and the second highest bid price, if any. Neill et al. [1994] found a significant difference in bids elicited using this “credibility feature.”

### *B. Dichotomous Choice Elicitation*

In response to Neill et al. [1994], many proponents of hypothetical surveys commented that this type of hypothetical bias was “well known” in open-ended elicitation procedures, and that it was precisely this type of unreliability which had prompted the use of dichotomous choice (DC) methods. A DC elicitation is just a “take it or leave it” offer, much like the posted-offer institution studied by experimental economists for many years. The difference is that the experimenter presents the subjects with a price, and the subject responds “yes” or “no” if she is willing to pay that amount. Incentive-compatibility is apparent, at least in the usual partial-equilibrium settings in which such things are discussed.

Cummings, Harrison and Rutström [1995] (CHR) designed some of the simplest experiments that have probably ever been run, to illustrate that DC designs were also subject to hypothetical bias. Subjects were randomly assigned to one of two rooms, the only difference being the use of hypothetical or real language in the instructions. An electric juicer was displayed, and passed around the room with the price tag removed or blacked-out. The display box for the juicer had some informative blurb about the product, as well as pictures of it “in action.” Subjects were asked to say whether or not they would be willing to pay some stated amount for the good.

Again, the hypothetical subjects responded much more positively than the real subjects. Since the private sources funding these experiments did not believe that “students were real people,” the subjects were non-student adults drawn from church groups. The same qualitative results were obtained with students, with the same commodity and with different commodities. Comparable results have been obtained in a WTA setting by Nape et al. [2003].

### C. Social Elicitation

There has been discussion in the literature that suggests that binary referenda are sufficient for incentive compatibility. However, as explained by Cummings, Elliott, Harrison and Murphy [1997; p.609ff.] (CEHM), a binary referendum is only necessary. To test if it was behaviorally sufficient, CEHM implemented simple majority rule experiments for an actual public good. After earning some income, in addition to their show-up fee, subjects were asked to vote on a proposition that would have each of them contribute a specified amount towards this public good. If the majority said “yes,” all had to pay. Again the key treatments were the use of hypothetical or real payments, and again there was significant evidence of hypothetical bias.<sup>16</sup>

One conceptual issue with the use of referenda to elicit valuations is whether the referenda should be viewed as a “framing metaphor” to help subjects comprehend the task, or as a “social choice microcosm” in which subjects actually make or influence choices for the population.<sup>17</sup> These alternatives imply quite different analyses of responses when an explicit “no vote” option is provided, and potentially different valuations. In the former case one would discard subjects that decide not to vote, but in the latter case that is part of the field social choice institution. The effect of these options deserves much more thought and experimental study.<sup>18</sup>

---

<sup>16</sup> Haab, Huang and Whitehead [1999] argue for allowing the residual variance of the statistical model to vary with the experimental treatment. They show that such heteroskedasticity corrections can lead the *coefficients* on the experimental treatment to become statistically insignificant, if one looks only at the coefficient of the treatment on the mean effect. This is true, but irrelevant for the determination of the *marginal effect* of the experimental treatment, which takes into account the joint effect of the experimental treatment variable on the mean response and on the residual variance. That marginal effect remains statistically significant in the original setting considered by Haab, Huang and Whitehead [1999], the referendum experiments of CEHM. Differences in standard errors of hypothetical and real responses are also important in the statistical calibration literature associated with pooling “stated choice” responses, discussed later.

<sup>17</sup> Art imitates life. Patterned after Brecht’s play, *The Resistable Rise of Arturo Ui*, the 1970 British farce *The Rise and Rise of Michael Rimmer* explored some of these ideas. The British population was encouraged to vote on all matters of national significance, and some of local sewerage significance, by interactive television survey every evening. The parallel to the manner in which some CVM surveys are now undertaken is eerie. The punchline is that after respondents are deluged for week after week with decisions, and grow tired of the task, the hero (Peter Cook) inserts a modest survey question making him the President of England for life as a substitute for the populace having to be burdened with these tasks while the pubs are still open. Co-written by John Cleese and Graham Chapman, of *Monty Python* fame, the plot has other parallels in contemporary lobbying efforts to have environmental valuation and punitive damage awards determined by “expert panels” rather than left to politicians, juries or the populace. So life may, in the end, imitate art.

<sup>18</sup> Carson et al. [1998] and Krosnick et al. [2002] undertake evaluations of this treatment using a replication of the hypothetical *Exxon Valdez* CVM survey.

#### *D. Multiple Price Lists and Revealed Preference Experiments*

Two “new valuation kids on the block” are simple extensions of the DC approach, to elicit choices from subjects over alternative configurations of goods that vary in terms of one or more attributes. In the simplest DC approach, the subject is given two choices: buy the good at a stated price or keep your money in the status quo.

In the first extension the experimenter implicitly offers the subject three choices: buy the good at one stated price, buy the good at another stated price, or keep your money. In this case, known in the experimental literature as a Multiple Price List (MPL) auction, the subject is actually asked to make two choices: say “yes” or “no” to whether the good would be purchased at the first price, and make a similar choice at the second price. The subject can effectively make the third choice by saying “no” to both of these two initial choices. The MPL can be made incentive-compatible by telling the subject that one of the choices will be picked at random for implementation.

The MPL design has been employed in three general areas in experimental economics: in the elicitation of risk attitudes by Holt and Laury [2002], in the elicitation of valuations for a commodity by Kahneman, Knetsch and Thaler [1990], and in the elicitation of individual discount rates by Coller and Williams [1999]. The use of the MPL has a longer history in the elicitation of hypothetical valuation responses, as discussed by Mitchell and Carson [1989; p. 100, fn. 14].

The MPL has three possible disadvantages. The first is that it only elicits interval responses, rather than “point” valuations. The second is that some subjects can switch back and forth from row to row, implying inconsistent preferences. The third is that it could be susceptible to framing effects, as subjects are drawn to the middle of the ordered table irrespective of their true values. Each of these potential problems can be addressed using appropriate designs and statistical procedures (e.g., Harrison et al. [2004]).

Coller and Williams [1999] provide test for hypothetical bias in the MPL format, and find that it exists. They show (p.121) that elicited discount rates are significantly higher in their hypothetical treatment, and exhibit a higher residual variance after correcting for differences in the demographic characteristics of

their samples.<sup>19</sup>

#### *E. Revealed Preference Experiments With Multiple Choices*

The other “new kid on the valuation block” involves several choices being posed to subjects, in the spirit of the revealed preference logic. Each choice involves the subject reporting a preference over two or more bundles, where a bundle is defined by a set of characteristics of one or more commodities. The simplest example would be where the commodity is the same in all bundles, but price is the only characteristic varied. This special case is just the MPL discussed above, in which the subject may be constrained to just pick one of the prices (if any). The most popular variant is where price and non-price characteristics are allowed to vary across the choices. For example, one bundle might be a lower quality version of the good at some lower price, one bundle might be a higher quality version at a higher price, and one bundle is the status quo in which nothing is purchased. The subject might be asked to pick one of these three bundles in one choice task (or to provide a ranking).

Typically there are several such choices. To continue the example, the qualities might be varied and/or the prices on offer varied. By asking the subject to make a series of such choices, and picking one at random for implementation, the subjects preferences over the characteristics can be “captured” in the familiar revealed preference manner. In an experimental setting respondents could be asked to make choices from each of several choice sets and one choice could be randomly drawn<sup>20</sup> as the binding set. The choice made by the individual would be implemented or realized. This method is obviously incentive-compatible. Furthermore, the incentive to reveal true preferences is relatively transparent.

This set of variants goes by far too many names in the literature. The expression “choice experiments” is popular, but too generic to be accurate. A reference to “conjoint analysis” helps differentiate the method, but at the cost of semantic opacity. In the end, the expression “revealed preference methods” serves to describe these methods well, and connect them to a long and honorable

---

<sup>19</sup> Holt and Laury [2002] also provide a test of hypothetical bias in the MPL format, and show that it also exists. Unfortunately, their design suffers from a simple confound: the comparable hypothetical and real responses are collected in a fixed order from the same subjects, so one cannot say whether it is the hypothetical consequences or the order that is generating differences in results.

<sup>20</sup> As a procedural matter, experimental economists generally rely on physical randomizing devices, such as die and bingo cages, when randomization plays a central role in the mechanism. There is a long tradition in psychology of subjects second-guessing computer-generated random numbers, and the unfortunate use of deception in many fields from which economists recruit subjects makes it impossible to rely on the subject trusting the experimenter in such things.

tradition in economics since Samuelson [1938], Afriat [1967] and Varian [1982][1983]. Moreover, it is inappropriate to have one term for a given elicitation method that uses hypothetical incentives and another term for the same method using real incentives.

Several studies examine hypothetical bias in this revealed preference elicitation method, at least as it is applied to valuation.

Carlsson and Martinsson [2001] allow subjects to allocate real money to 2 environmental projects, varying 3 characteristics: the amount of money the subject personally receives, the amount of money donated to an environmental project by the researchers, and the specific World Wildlife Fund project that the donation should go to. They conclude that the real and hypothetical response are statistically indistinguishable, using statistical models commonly used in this literature. Several problems with their experiment make it hard to draw reliable inferences. First, and most seriously, the real treatments were all in-sample: each subject gave a series of hypothetical responses, and then gave real responses. There are obvious ways to test for order effects in such designs, as used by CHR for example, but they are an obvious confound here. Directly comparable experiments by Svedsäter and Johansson-Stenman [2001] suggest that order effects were in fact a significant confound. Second, the subjects were allocating “house money” with respect to the donation, rather than their own. This made it hard to implement a status quo decision, since it would have been dominated by the donation options if the subject had even the slightest value for the environmental project. On the other hand, there is a concern that these are all artificial, forced decisions that might not reflect how subjects allocate monies according to their true preferences (unless one makes strong separability assumptions). Third, all three environmental projects were administered by the same organization, which leads the subject to view them as perfect substitutes. This perception is enhanced by a (rational) belief that the organization was free to re-allocate un-tied funds residually, such that there is no net effect on the specific project. Thus the subjects may well have rationally been indifferent over this characteristic.<sup>21</sup>

Lusk and Schroeder [2004] conduct a careful test of hypothetical bias for the valuation of beef using revealed preference methods. They consider 5 different types of steak, and vary the relative prices of

---

<sup>21</sup> When subjects are indifferent over options, it does not follow that they will choose at random. They might use other heuristics to pick choices which exhibit systematic biases. For example, concern with a possible left-right bias leads experimental economists looking at lottery choice behavior to randomize the order of presentation.

each steak type over 17 choices. For the subjects facing a real task, one of the 17 choices was to be selected at random for implementation. Subjects also considered a “none of these option” that allowed them not to purchase any steak. Each steak type was a 12oz steak, and subjects were told that the baseline steak, a “generic steak” with no label, had a market price of \$6.07 at a local supermarket. Each subject received a \$40 endowment at the outset of the experiment, making payment feasible for those in the real treatment. Applying the statistical methods commonly used to analyze these data, they find significant differences between hypothetical and real responses. Specifically, they find that the marginal values of the attributes between hypothetical and real are identical but that the propensity to purchase, attributes held constant, is higher in the hypothetical case.

More experimental tests of the revealed preference approach are likely. I conjecture that the experimental and statistical treatment of the “no buy” option will be important to the evaluation of this approach. It is plausible that hypothetical bias will manifest itself in the “buy something” versus “buy nothing” stage in decision-making, and not so much in the “buy this” or “buy that” stage that conditionally follows. Indeed, this hypothesis has been one of the implicit attractions of the method. The idea is that one can then focus on the second stage to ascertain the value placed on characteristics. But this promise may be illusory if one of the characteristics varied is price, or if separability in decisions is not appropriate.

This point is of more importance for the use of revealed preference methods for public goods valuation than for their traditional marketing applications. The latter tends to focus on attribute values because most firms are keenly interested in product differentiation issues with new or modified brands. The former is almost exclusively focused on the “buy” or “not buy” margin of choice. Thus it is understandable that evidence of significant hypothetical bias in the “buy” or “not buy” stage is of less concern to those trying to place values on alternative characteristics sets, but is central for valuations of public goods.

## **2. Instrument Calibration**

Much of the debate and controversy over “specifications” in the CVM literature concerns the choice of words. The problem of “choosing the right words” in CVM studies has assumed some

importance through the result of judicial decisions. In 1989 the U.S. District Court of Appeals, in *State of Ohio v. U.S. Department of the Interior* (880 F.2nd. at 474), asserted that the “... simple and obvious safeguard against overstatement [of WTP], however, is more sophisticated questioning.” (p.497). While disagreeing that this process is “simple and obvious,” it is apparent that one can only assess the improvement from different CV questionnaires if one has a way of knowing if any bias is being reduced. This mandates the use of some measure of the real economic commitment that a subject would make in the same setting as the hypothetical question. The laboratory is clearly one place where such measures can be readily generated.

CVM research partially addresses these “wording issues” by employing “focus groups” to help guide the initial survey design and/or employing variations in the final survey design. With design variation, if survey results are similar across several different versions of a questionnaire, then there is some presumption that the *hypothetical* responses do not depend on the particular words chosen from this set. If the results are not similar across survey design, then they provide some bounds on the hypothetical response. See, for example, Imber, Stevenson and Wilks [1991] and Rowe, Schulze, Shaw, Schenk and Chestnut [1991].

Of course, there is no claim in these studies to have shown that any of these versions is any closer than the other to the real economic commitments that subjects would make. The only claim is that they all might give comparable hypothetical numbers or bounds on the hypothetical WTP. A pack of drunk rednecks agreeing that the UCF Golden Knights are the best college football team does not, sadly, make it so.

The increasing use of “focus groups” in CVM research, in which subjects are directly asked to discuss their interpretation of CV questions (e.g., see Smith [1990][1992] or Schkade and Payne [1993]), is a practical response to the concerns cognitive psychologists have long expressed about the importance of different ways of presenting valuation questions. For example, Fischhoff and Furby [1988] and Fischhoff [1991] correctly characterize many possible disparities between the way subjects perceive the CVM valuation question and the way investigators perceive it. Useful as “focus groups” might be to help avoid misinterpretation of a survey instrument in some manner, are “focus groups” to be judged effective in reducing bias? In the absence of comparable responses from real behavior, one can only speculate on the importance of what is learned from “focus groups” in demonstrating that any alternative wordings for a

survey may reduce (or exacerbate) hypothetical bias.

The laboratory provides a simple metric by which one can test, in meaningful ways, the importance of different presentations of valuation questions. Because controlled laboratory experiments may be used to enforce real economic commitments, they provide “benchmarks” to which alternative scenario designs, or wording choices, may be evaluated in their effectiveness of reducing hypothetical bias. Thus, using laboratory experiments is likely to be more informative than the casual introspective nature of the literature on wording choice in survey design.<sup>22</sup> The problem of deciding which set of words is “best” might, in some instances,<sup>23</sup> be easily and directly tested using controlled laboratory experiments such as those presented earlier.

The idea of instrument calibration has already generated two important innovations in the way in which hypothetical questions have been posed: recognition of some uncertainty in the subject’s understanding of what a “hypothetical yes” means, and the role of “cheap talk” scripts directly encouraging subjects to avoid hypothetical bias.

### **2.1 Semantic Ambiguity When “No Means No, But Yes May Also Mean No”**

One of the hypotheses that Cummings, Harrison and Rutström [1995; fn.5, p.261] advanced to explain their results was that the subjects might have just misunderstood the question, simple as it might seem. It is plausible that subjects in the hypothetical treatments might have interpreted the subjunctive language as asking in effect, “would you *ever* be willing to pay this amount of money for this juicer?” Such subjects might then begin a long search over possible states of nature under which they would be led to buy the good at this price, having no regard to either the current state of nature or the probabilities of alternative states occurring.<sup>24</sup> The subjects in the real treatment, however, were encouraged by the “reality check” of having to open their wallet to answer the question, “would you be willing to pay this amount of money for this juicer *here and now*?” The differences between the treatments could be just due to the fact

---

<sup>22</sup> This is not to disavow the use of casual introspection, particularly when it is prohibitively costly in terms of money or time to collect data in the laboratory. Many of the wording and logistical suggestions of Dillman [1978], for example, seem plausible and sensible enough *a priori* that one would not bother applying scarce research dollars testing them.

<sup>23</sup> The qualification here refers to the existence of a feasible and affordable laboratory procedure for eliciting valuations from subjects that are truthful. Such procedures clearly exist for private, deliverable goods. They also exist for public, deliverable goods under certain circumstances. They do not presently exist for non-deliverable goods.

<sup>24</sup> Another interpretation might be “what is this juicer worth?” which is again different from the question that the experimenter wanted an answer to.

that they interpreted the valuation questions differently.

However, it is precisely this type of interpretational problem which could be at the heart of the observed hypothetical bias. The questions listed above were not the ones asked by the experimenters; apart from the use of subjunctive terms in the hypothetical treatment, the questions actually asked were identical. Hence this is not a flaw in experimental design, but points to a promising avenue for constructive attempts to mitigate hypothetical bias in general.

Blumenschein et al. [1998] provide just such an attempt, in a replication of the CHR experiments. Apart from some wording changes to try to make the hypothetical subjects aware that they are being asked if they would buy the good here and now, they followed-up all hypothetical “yes” responses by asking subjects to state if they were “fairly sure” or “absolutely sure” they would buy the good. By taking the latter responses *only* as indicating a “yes,” they conclude that hypothetical bias disappears. Although this is a simple device for eliciting the likelihood that a subject would actually follow through with the planned action, it appears to deliver some improvements in the ability of *this* hypothetical question to match *this* real question. The obvious issue to be resolved with further experiments is the generality of this result (e.g., see Blumenschein et al. [2001]). Even without further tests, however, this extension seems obviously sensible, providing the words are explained openly to the subjects.

## 2.2 “Cheap Talk”

Another application of the notion of instrument calibration using lab experiments was provided in a series of experiments, reported in Cummings, Harrison and Osborne [1995a], which evaluated two sets of “wording” issues.<sup>25</sup> Various “cheap talk” script is introduced into the survey design of CEHM, and the new results are compared to their baseline (i.e., no “cheap talk”) results. In the theory of games, “cheap talk” games are those in which game participants may send signals to each other costlessly, and where those signals do not change the payoffs from adopting alternative strategies. The notion of a “cheap talk”

---

<sup>25</sup> Subsequent experimental investigations of the effect of cheap talk include Cummings and Osborne [1999] and List [2001], Aadland and Caplan [2003], Brown, Ajzen and Hrubec [2003] and Murphy, Stevens and Weatherhead [2003]. There have also been some new applications of the idea with respect to reminders of budget sets and substitutes, although these studies have been limited to the effect on one hypothetical survey compared to another (e.g., Loomis, Gonzales-Caban and Gregory [1994]). Such exercises have some value, since we know that if they do not change the response at all then they cannot be changing the response in the direction of mitigating presumed hypothetical bias. But to measure if bias is reduced significantly one cannot do without the metric of a real response that is incentive compatible.

script can be defined analogously. It is the additional verbiage (and thus signals) that the experimenter may send to the experiment participants, and these signals are costless in terms of the experiment design.

Two extensions to the language used in the CEHM experiments that introduced subjects to the hypothetical referendum are considered. The first is referred to as “Light Cheap Talk” and the second, imaginatively enough, as “Heavy Cheap Talk.” The latter involves an extension of the ideas expressed in the former, so one can think of the Light version as being nested in the Heavy version.

Both “cheap talk” scripts respond to the current CVM protocol that surveys remind respondents of the opportunity costs their choices could imply if the survey were in fact eliciting real economic commitments. This protocol is emphasized by the National Oceanic and Atmospheric Administration’s (NOAA) proposed rules for the conduct of CV surveys for natural resource damage assessment, which require that (NOAA [1994b; p.23101]):

Prior to the value elicitation [in CV surveys], respondents shall be reminded of their budget constraints and their alternative expenditures. Respondents shall be reminded that their WTP for the environmental program in question would reduce their expenditures on other goods. This reminder should be more than perfunctory, but less than overwhelming.

The proposed rules in effect advocate the use of “cheap talk” to cajole respondents into searching their preferences as if they faced a “real” economic commitment implying opportunity costs.

Both experiments reported in Cummings, Harrison and Osborne [1995a] have in common an approach to using “cheap talk” as a tool to reduce hypothetical bias that is different than that advocated by current CVM protocol. Rather than just reminding respondents of budget constraints and/or substitute commodities, they chose to inform respondents directly of the issue of hypothetical bias. In both versions of the “cheap talk” script, the respondents were given an introduction to hypothetical bias as an issue researchers have been concerned with in hypothetical surveys/referenda such as the one in which the respondents were participating.

The “Heavy Cheap Talk” version significantly expands the description of hypothetical bias and presents the respondents with the hypothetical bias results from CEHM. The “Heavy Cheap Talk” script also includes several paragraphs of text which conforms with NOAA’s recommendation that respondents be reminded of the potential opportunity costs their choice would incur if the survey actually required an economic commitment.

In both the “Light Cheap Talk” and “Heavy Cheap Talk” experiments, the additional script

appeared just prior to the referendum vote. In addition to reading the “cheap talk” scripts to the subjects, the scripts were provided to respondents as written inserts in their respondent packets, which they were instructed to open only at the appropriate time.<sup>26</sup>

Using the comparable hypothetical and real<sup>27</sup> referendum experiments of CEHM as the control, and deleting a handful of subjects with missing observations for some key variables, there are 354 individual observations. Of these, 29 participated solely in the “Light Cheap Talk” experiments and 50 in the “Heavy Cheap Talk” experiments, implying about 22% of the sample were given some version of the additional verbiage.

The results are reported in the form of a binomial probit model. The treatment variables are the only factors which are statistically significant predictors of voting behavior when taken individually, but all factors taken together are highly significant (the regression as a whole is significant at a 5.4% critical value). The “Light Cheap Talk” script has the opposite effect of the “Heavy” and “Real” treatments. The “Light” treatment *increases* the probability of a Yes response by about 0.21, whereas the “Heavy” and “Real” treatments *decrease* it by virtually the same amount (0.22 and 0.23, respectively). These results indicate that the “Light” treatment increases hypothetical bias, while the “Heavy” treatment reduces hypothetical bias.<sup>28</sup>

To the extent that the “Light Cheap Talk” highlighted to respondents the hypothetical nature of the referendum we might expect to see increases in the numbers who vote “Yes.” In the hypothetical referenda with no “cheap talk,” some respondents may doubt that the referenda would indeed be hypothetical after the vote. Adding the “cheap talk” script could assure those who might have doubted the no “cheap talk” scenario as not being hypothetical.

The second result from these experiments is the success in finding a form of “cheap talk” that statistically eliminated hypothetical bias. The “Heavy Cheap Talk” version of the script resulted in the same reduction in the probability of a “yes” vote as the “Real” referenda, when both are compared to the

---

<sup>26</sup> Providing script in writing for respondents to read during the experimenter’s verbal discourse is a commonly used method to focus respondent attentions on the content of the script.

<sup>27</sup> References to the experiments conducted by CEHM in which an actual payment was made by subjects if the referendum passed will be referred to as the “real” treatment, or “real” version of the script.

<sup>28</sup> Wald tests were conducted to compare the effect of these referenda treatments. We reject the null hypothesis that the effect of “Light Cheap Talk” on the probability that a respondent votes Yes is the same as the effect of either the “Heavy Cheap Talk” or the “Real” referenda (at the 95.67% and 99.88% level of significance, respectively). However, we *cannot* reject the null hypothesis that the “Heavy Cheap Talk” and “Real” referenda have the same effect on voting behavior (the  $\chi^2$  statistic for this test is significant at only the 6.20% level of significance).

hypothetical survey with no “cheap talk.” It is important to emphasize that this is just a behavioral existence proof. The lab experiments show in this *instance* that there *exist* one set of words which will result in a hypothetical survey generating the same results as a comparable survey eliciting real economic commitments. To give the notion of cheap talk its best chance of succeeding, this design exploited pre-existing knowledge about the extent of hypothetical bias in a control experiment (CEHM). It does not follow that these methods can be generally expected to be successful in reducing hypothetical bias in referendum experiments, although we are entitled to some degree of optimism. A particularly important extension of this line of research will be to see if more generic scripts will have the same effect. That is, will we get the same effects if we talk about hypothetical bias in general, abstract terms, rather than display precise quantitative results for a directly comparable subject pool and issue?

Moreover, it is arguable that the Heavy Cheap Talk script is not cheap talk at all, and is in fact confounded with a change in the very commodity being valued. A key feature of the notion of cheap talk is that it must not change the expected payoffs to subjects from given strategies. If the cheap talk script changes the subject’s perception of the quality of the good, or the market value of the good, then it is not cheap in this sense: it is changing the attributes of the good being valued. It is at least arguable that the Heavy Cheap Talk scripts do this, since they are in effect telling subjects that the product is “actually worth more than most of you typically believe.”

Cummings and Taylor [1999] report the results of the Heavy Cheap Talk treatment, and one modification that removes the precise numerical information but leaves in much of the language exhorting the respondent not to suffer from hypothetical bias. Apart from the removal of the text documenting the numerical extent of the hypothetical bias in the control experiment, the text is virtually identical to the Heavy Cheap Talk treatment. They conclude on an optimistic note:

the cheap talk design was successful in eliciting responses to hypothetical valuation questions that were indistinguishable from responses to valuation questions involving actual payments. This finding was robust across changes in the cheap talk script and changes in the experimental design that may be of importance for field applications of the method. We have interpreted these results as suggesting that the cheap talk design can be effective in eliminating hypothetical bias.

They do mention (p.659, fn.18) that there “are of course any number of other changes in the cheap talk script that might warrant examination. See Cummings et al. (1995a) for a discussion of some of our earlier pretests with shorter versions of the cheap talk script.” It is a pity that they did not mention that the Light

Cheap Talk script actually *worsened* hypothetical bias, since that is such a strikingly negative result.<sup>29</sup> The reason that this negative result is so important is that instrument calibration in general, and cheap talk in particular, should never be viewed as a “magic bullet” to remove hypothetical bias.

The optimistic conclusions of Cummings and Taylor [1999] provided an easy target for List [2001], Aadland and Caplan [2003] and Brown, Ajzen and Hrubes [2003], who all show that cheap talk scripts do not work for all subject groups.

### **2.3 The Seductive Hint of Realistic Consequences**

One feature of hypothetical surveys in the field is not well captured by these experiments: the chance that the subject’s hypothetical response might influence policy or the level of damages in a lawsuit. To the extent that we are dealing with a subjective belief, such things are intrinsically difficult to control perfectly. In some field surveys, however, there is a deliberate use of explicit language which invites the subject to view their responses as having some chance of affecting real decisions.

If one accepts that field surveys are successful in encouraging *some* subjects to take the survey for real in a subjectively probabilistic sense, then the natural question to ask is: “how realistic does the survey have to be, in the eyes of respondents, before they respond *as if* it were actually real?” In other words, if one can encourage respondents to think that there is some chance that their responses will have an impact, at what point do the subjects behave the way they do in a completely real survey? Obviously this question is well-posed, since we know by construction that they must do so when the chance of the survey being real is 100%. The interesting empirical question, which we examine, is whether any smaller chance of the survey being real will suffice. This question takes on some significance if one can show that the subject will respond realistically even when the chance of the payment and provision being real is small.

#### *A. Field Counterparts*

Many field surveys are designed to avoid the problem of hypothetical bias. Great care is often taken in the selection of motivational words in cover letters, opening survey questions, and key valuation questions, to encourage the subject to take the survey seriously in the sense that their response will

---

<sup>29</sup> Cummings and Taylor [1999; p.664] only note that efforts “to obtain unbiased valuation responses with a much shorter version of the cheap talk script were unsuccessful.”

“count”. It is not difficult to find many prominent examples of this pattern.

Consider the generic cover letter advocated by Dillman [1978; pp.165ff.] for use in mail surveys. The first paragraph is intended to convey something about the social usefulness of the study: that there is some policy issue which the study is attempting to inform. The second paragraph is intended to convince the recipient of their importance to the study. The idea here is to explain that their name has been selected as one of a small sample, and that for the sample to be representative they need to respond. The goal is clearly to put some polite pressure on the subject to make sure that their socio-economic characteristic set is represented.

The third paragraph ensures confidentiality, so that the subject can ignore any possible repercussion from responding one way or the other in a “politically incorrect” manner. Although seemingly mundane, this assurance can be important when the researcher interpretes the subject as responding to the question at hand rather than uncontrolled perceptions of repercussions. It also serves to mimic the anonymity of the ballot box.

The fourth paragraph builds on the preceding three to drive home the usefulness of the survey response itself, and the possibility that it will influence behavior:

The fourth paragraph of our cover letter reemphasizes the basic justification for the study -- its social usefulness. A somewhat different approach is taken here, however, in that the intent of the researcher to carry through on any promises that are made, often the weakest link in making study results useful, is emphasized. In {an example cover letter in the text} the promise (later carried out) was made to provide results to government officials, consistent with the lead paragraph, which included a reference to bills being considered in the State Legislature and Congress. Our basic concern here is to make the promise of action consistent with the original social utility appeal. In surveys of particular communities, a promise is often made to provide results to the local media and city officials. (Dillman [1978; p.171])

From our perspective, the clear intent and effect of these admonitions is to attempt to convince the subject that their response will have some probabilistic bearing on actual outcomes. We do not need to enter into any debate on whether this intent is realized.

This generic approach has been used, for example, in the CVM study of the *Nestucca* oil spill by Rowe, Schulze, Shaw, Schenk and Chestnut [1991]. Their cover letter contained the following sentences in the opening and penultimate paragraphs:

Government and industry officials throughout the Pacific Northwest are evaluating programs to prevent oil spills in this area. Before making decisions that *may cost you money*, these officials want your input. [...] The results of this study will be made available to representatives of state, provincial and federal governments, and industry in the Pacific Northwest. (emphasis added).

In the key valuation question, subjects are motivated by the following words:

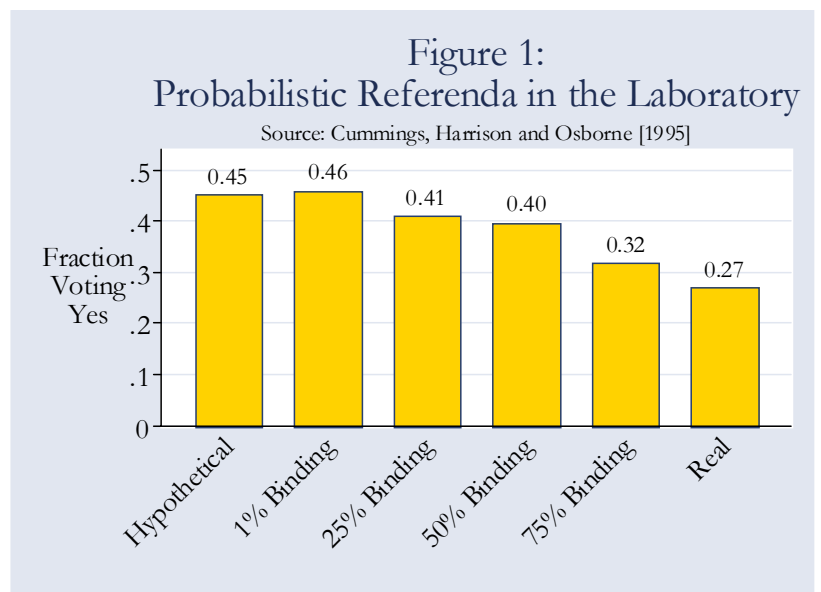
Your answers to the next questions are very important. We do not yet know how much it will cost to prevent oil spills. However, to make decisions about new oil spill prevention programs that could cost you money, government and industry representatives want to learn how much it is worth to people like you to avoid more spills.

These words reinforce the basic message of the cover letter: there is some probability, however small, that the response of the subject will have an actual impact.

We conclude from this example that *even if* field survey researchers would be willing to accept the claim that “hypothetical surveys are biased in relation to real surveys,” they might deny that they actually conduct hypothetical surveys. Without entering into a debate about how realistic the surveys are as the result of direct or implied “social usefulness,” their claim must be that a little bit of reality elicits the same responses as The Real Thing.

### B. Experimental Results

Cummings, Harrison and Osborne [1995b] and Cummings and Taylor [1998] report striking results with this design, illustrated in Figure 1. As the probability of the real economic commitment being binding increased from 0% to 1%, there was virtually no change in the fraction voting yes. In fact, it went up from 45% to 46%, and while that is not a statistically significant increase it is reminiscent of the effect of Light Cheap Talk: plausibly, the information that there was only a 1% chance that the referendum was to be binding served to remind some respondents that it was not real. As the chance of the referendum response being binding increases to 25%, 50% and 75%, the responses get closer and closer to those obtained when it was completely binding.



It remains an open question how subjects in the field might interpret the advisory nature of their responses. If a sample of 100 was contacted and told that their responses would be scaled up to the whole population and applied without fail, would that be interpreted by the subject as a 1% chance of the referendum response being binding or as a 100% chance if it being binding? The latter would be a sensible interpretation, but then the respondent must decide how likely it is that their response will be pivotal. And then the respondent needs to evaluate the chances of this survey sample being binding, given the nature of the political and litigation process. These issues of perception deserve further study. However, the results to hand do not suggest that by making the probability of the response binding by “epsilon” that responses will be exactly the same as if they were 100% binding.

List, Berrens, Bohara and Kerkvliet [2004] examine these issues by varying the “social isolation” of the subject in hypothetical and real settings. Their default elicitation procedure involves a simple majority rule referendum in which subjects vote for each person to contribute \$20 towards an environmental research center. They undertake hypothetical and real versions of this default, following CEHM. They then consider two variants. In one case they have some subjects, picked at random, announce their vote in public; in another variant they use a “randomized response” procedure to ensure that votes are private. The latter variant is of some significance since it corresponds to the secrecy provided by the ballot box in the field.<sup>30</sup> Unfortunately, there is a fundamental logical difficulty applying randomized response methods to a social elicitation task: there is no way for the experimenter to know if a “yes” vote is for the project in question or because the response is given to an unrelated question. One can make statistical inferences about the true response to the project in question, particularly if the randomizing device has known

---

<sup>30</sup> List et al. [2004; p.749] draw some sweeping conclusions from their results: “... our findings raise serious concerns about the experimental results in the literature purporting to measure hypothetical bias given the specific social context in which some of the studies have been conducted. For example, are ‘actual’ statements of value in these experiments providing accurate signals of true preferences? And, what is the correct benchmark if the degree of social isolation is not controlled? While our study only pertains directly to the referendum voting institution, our findings raise the specter that social isolation effects may be important in every elicitation format, including openended valuation questions, choice experiments, dichotomous choice questions, etc.” These concerns are unfounded. Previous experiments used one particular “social context,” and did not leave this uncontrolled at all. The semantic distinction between “true preferences” and “actual context-specific preferences” is surely contrived: it is better to assume that all preferences are context-specific unless theory or evidence suggests otherwise. The fundamental issue here is whether “social context” *interacts* with the use of hypothetical or real consequences. Existing experiments have assumed the absence of such an interaction, which is worthy of study.

probabilities of being applied. But that is a very different thing than knowing with certainty if 50% or more voted to have \$20 taken away from them. The relevant instructions included these lines:

If more than 50% of you vote “yes” on this proposition, all of you will pay \$20.00. [...] If 50% or fewer of you vote “yes” on this proposition, no one will pay \$20.00, we will not send a check to the center and the start-up expenses will not be gathered. We are now passing out a ballot. Remember how the vote works. If more than 50% vote “yes” we will collect \$20.00 from each of you, and we will mail this check to the center right here today. If 50% or less vote “yes,” no one will pay \$20.00, and we will not mail this check to the center. Any questions? To ensure your privacy, we are using a technique in which you are asked to give a truthful response to a sensitive question or answer a trivial question of fact. So that only you will know which question you answered, it will first be necessary that you compute a random number from your Social Security Number. [...] Please answer the question honestly. Your answer cannot be traced to you, nor do we have any interest in doing so. We are only interested in trying a statistical procedure. If your random number is between 0 and 10, answer question A below. If your random number is between 11 and 36, answer question B below. (A) Is your mother’s birthday in May or June? (B) Are you are willing to pay \$20 to support the creation of [the research center]?

At this point the astute subject should have asked for clarification, since these instructions are incoherent: if the answer cannot be traced to question A or B, then there is no way for the experimenter to know if 50% voted for the research center. One cannot say what was motivating the subject in situations such as this. If the subject wanted to avoid this ambiguity, and disliked the risk of imposing a \$20 tax on those actually voting “yes” to question A, they might vote “no” when their true intention was to support the research center.

### 3. Statistical Calibration

#### 3.1 The Basic Idea

Can a decision maker *calibrate* the responses obtained by a hypothetical survey so that they more closely match the real economic commitments that the subjects would have been expected to make? A constructive answer to this question was offered by Blackburn, Harrison and Rutström [1994]. The essential idea underlying this approach is that the hypothetical survey provides an informative, but statistically biased, indicator of the subject’s true willingness to pay for the environmental good. The trick is how to estimate and apply such bias functions.<sup>31</sup> They propose doing so with the *complementary* use of

---

<sup>31</sup> The possible use of estimated bias functions for public goods valuation was first proposed by Kurz [1974]. A subsequent, and brief, discussion of the idea appears in Freeman [1979]. Although restricted to the private goods experiments of CHR, Blackburn, Harrison and Rutström [1994] appears to be the first application and test of the idea. Similar ideas motivate the econometric approach advocated by McClelland, Schulze, Waldman, Irwin and Schenk [1991]. Finally, the idea of bias function estimation was raised by Roy Radner at a public meeting into the use of Contingent Valuation Method conducted

field elicitation procedures that use hypothetical surveys, laboratory elicitation procedures that use hypothetical and non-hypothetical surveys, and laboratory elicitation procedures that use incentive-compatible institutions.<sup>32</sup>

The upshot of the statistical calibration approach is a simple comparison of the original responses to the hypothetical survey and a set of calibrated responses that the same subjects *would have made* if asked to make a real economic commitment in the context of an incentive-compatible procedure. This approach does not predetermine the conclusion that the hypothetical survey is “wrong.” If the hypothetical survey is actually eliciting what its proponents say that it is, then the calibration procedure should say so. In this sense, calibration can be seen as a way of validating “good hypothetical surveys” and correcting for the biases of “bad hypothetical surveys.”

The statistical calibration approach can do more than simply pointing out the possible bias of a hypothetical survey. It can also evaluate the confidence with which one can infer statistics such as the population mean from a given survey. In other words, a decision maker is often interested in the bounds for a damage assessment that fall within prescribed confidence intervals. Existing hypothetical surveys often convey a false sense of accuracy in this respect. A calibration approach might indicate that the population mean inferred from a hypothetical survey is reliable in the sense of being unbiased, but that the standard deviation was much larger than the hypothetical survey would directly suggest. This type of extra information can be valuable to a risk-averse decision maker.

Consider the analogy of a watch that is always 10 minutes slow to introduce the idea of a *statistical bias function* for hypothetical surveys. The point of the analogy is that hypothetical responses can still be informative about real responses if the bias between the two is systematic and predictable. The watch that is always 10 minutes slow can be informative, but only if the error is *known* to the decision maker and if it

---

under the auspices of the National Oceanic and Atmospheric Administration [1992]. He asked one speaker, “... what would be a practical method, if any, of taking the results of the CVM willingness to pay and adjusting them ... in order to come to a damage assessment? How would one go about that?” (p. 99), and later followed up with a related question: “... are there things that one can do when one does the CVM, if one were to do it, and that would minimize this bias and, secondly, enable one to estimate it?” (p. 100). The studies reviewed here attempt to provide answers to these questions.

<sup>32</sup> Related work on statistical calibration functions includes Fox et al. [1994], Johannesson et al. [1999], Harrison et al. [1999] and List and Shogren [1998][2002]. Some of this is discussed further in §3.3 below.

is *transferable* to other instances (i.e., the watch does not get further behind the times over time).

Blackburn, Harrison and Rutström [1994] define a “known bias function” as one that is a systematic statistical function of the socio-economic characteristics of the sample. If this bias is not mere noise then one can say that it is “knowable” to a decision maker. They then test if the bias function is transferable to a distinct sample valuing a distinct good, and conclude that it is. In other words, they show that one can use the bias function estimated from one instance to calibrate the hypothetical responses in another instance, and that the *calibrated hypothetical* responses statistically match those observed in a paired *real* elicitation procedure. Johannesson et al. [1999] extend this analysis to consider responses in which subjects report the confidence with which they would hypothetically purchase the good at the stated price, and find that information on that confidence is a valuable predictor of hypothetical bias.

How do we anticipate these calibration rules working in practice? Subjects will be recruited to participate in a laboratory DC valuation exercise in which economic commitments are real. The commodity used in this real valuation exercise, the *surrogate good*, will be a private, deliverable good which shares many of the attributes of the *target good*, the commodity of primary interest. The target good will presumably be the subject of a larger-scale hypothetical CVM survey. In some settings we could imagine that the surrogate and target goods are identical, such as when one undertakes a marketing survey for a new good. Actual prototypes of the new good are typically developed for the purpose of estimating potential demand. Such prototypes may be used in a real DC laboratory valuation exercise as well as in a large-scale, hypothetical CVM exercise (e.g., see Hoffman et al. [1993] and Hayes et al. [1995]).

Given the observed real responses collected in the laboratory we may estimate a *real* valuation function. Given this estimated real valuation function, and a vector of demographic characteristics for the hypothetical sample, we can predict the real response probabilities of the subjects who participated in the large-scale CVM (e.g., see Harrison and Lesley [1996]). Similarly, the observed hypothetical responses from the national CVM may be used to obtain an estimate of a *hypothetical* valuation function and to predict hypothetical response probabilities.

Given the observed hypothetical responses, the predicted hypothetical response probabilities based

on the CVM valuation function and the predicted real response probabilities based upon the real laboratory valuation function, we can calibrate the hypothetical CVM responses. Calibration rules can be used for assessing the probability that a subject will respond “yes” in a real treatment given that she has responded “yes” in a hypothetical treatment, and calibration rules can similarly be used for assessing this probability given that a subject has responded “no” in a hypothetical scenario. Thus, the laboratory valuation exercise is used as a *complement* to the large-scale CVM valuation exercise in the field.

The more likely, and more controversial, setting is where the target and surrogate goods are distinct. Such a setting is more likely given that the primary scope of applications of hypothetical survey methods, such as the CVM, is to elicit valuations for commodities that are not deliverable in any cost-effective or credible manner. This setting is more controversial since it will be a matter of judgement if the chosen surrogate good adequately represents the characteristics of the target good. Nevertheless, in the absence of valuation mechanisms that can elicit values for non-deliverable goods which are demonstrably reliable in the sense that they accurately reflect a “true” value, such an indirect approach is a feasible alternative.

### **3.2 Pooling Responses From Different Mechanisms**

Building on long-standing approaches in marketing, a different statistical calibration tradition seeks to recover similarities and differences in preferences from data drawn from various institutions. The original objective was “data enrichment,” which is a useful way to view the goal of complementing data from one source with information from another source.<sup>33</sup> Indeed, the exercise was always preceded by a careful examination of precisely what one could learn from one data source that could not be learned from another, and those insights were often built into the design. For example, attribute effects tend to be positively correlated in real life: the good fishing holes have many of the positive attributes fishermen want. This makes it hard to tease apart the effects of different attributes, which may be important for

---

<sup>33</sup> See Hensher, Louviere and Swait [1999] and Louviere, Hensher and Swait [2000; chs.8, 13] for reviews.

policy evaluation. Adroit combination of survey methods can mitigate such problems, as illustrated by Adamowicz, Louviere and Williams [1994].

Relatively few applications of this method have employed laboratory data, such that there is at least one data generating mechanism with known incentive compatibility. One exception is Cameron, Poe, Ethier and Schulze [2002]. They implement 6 different hypothetical surveys, and one actual DC survey. All but one of the hypothetical surveys considered the same environmental good as the actual DC survey; the final hypothetical survey used a “conjoint analysis” approach to identify attributes of the good. Their statistical goal was to see if they could recover the same preferences from each data generation mechanism, with allowances for statistical differences necessitated by the nature of the separate responses (e.g., some were binary, and some were open-ended). They develop a mixture model, in which each data generation mechanism contributes to the overall likelihood function defined over the latent valuation. Although they conclude that they were generally able to recover the same preferences from most of the elicitation methods, their results depend strikingly on the assumed functional forms.<sup>34</sup> Their actual DC response was only at one price, so the corresponding latent WTP function can only be identified if one is prepared to extrapolate from the hypothetical responses. The upshot is a WTP function for the actual response that has a huge standard error, making it hard to reject the null that it is the “same” as the other WTP functions. The problems are clear when one recognizes that the only direct information obtained is that only 27% of the sample would purchase the environmental good at \$6 when asked for real, whereas 45% would purchase the good when asked hypothetically.<sup>35</sup> The only information linking the latent WTP functions is the reported income of respondents, along with a raft of assumptions about functional form.

A popular approach to combining data from different sources has been proposed in the stated choice literature: see Louviere, Hensher and Swait [2000; ch. 8, 13] for a review. One concern with this approach is that it relies on differences in an unidentified “scale parameter” to implement the calibration.

---

<sup>34</sup> Unfortunately the data from this study are not available for public evaluation because of confidentiality agreements with the public utility involved in the field study (Trudy Cameron; personal communication), so one cannot independently assess the effects of alternative specifications.

<sup>35</sup> This compares the 0-ACT and 1-PDC treatments, which are as close as possible other than the hypothetical nature of the response elicited.

Consider the standard probit model of binary choice, to illustrate. One common interpretation of this model is that it reflects a latent and random utility process in which the individual has some cardinal number for each alternative that can be used to rank alternatives. This latent process is assumed to be composed of a deterministic core and an idiosyncratic error. The “error story” varies from literature to literature,<sup>36</sup> but if one further assumes that it is normally distributed with zero mean *and unit variance* then one obtains the standard probit specification in which the likelihood contribution of each binary choice observation is the cumulative distribution function of a standard normal random variable evaluated at the deterministic component of the latent process. Rescaling the assumed variance only scales up or down the estimated coefficients, since the contribution to the likelihood function depends only on the cumulative distribution below the deterministic component. In the logit specification a comparable normalization is used, in which the variance is set to  $\pi^2/3$ . Much of the “data enrichment” literature in marketing assumes that the two data sources have the same deterministic component, but allows the scale parameter to vary. This has nothing to say about calibration, as conceived here.

But an extension of this approach does consider the problem of testing if the deterministic components of the two data sources differ, and this nominally has more to do with calibration. The methods employed here were first proposed by Swait and Louviere [1993], and are discussed in Louviere, Hensher and Swait [2000; §8.4]. They entail estimation of a model based solely on hypothetical responses, and then a separate estimation based solely on real responses. In each case the coefficients on the explanatory variables (e.g., sex, age) conditioning the latent process are allowed to differ, including the intercept on the latent process. Then they propose estimation of a “pooled” model in which there is a dummy variable for the data source. Implicitly the pooled model assumes that the coefficients on the explanatory variables *other than the intercept* are the same for the two data sources.<sup>37</sup> The intercepts implicitly differ, if one thinks of there being one latent process for the hypothetical data and one latent process for

---

<sup>36</sup> In the stated choice literature they refer to unobserved individual idiosyncracies of tastes (e.g., Louviere, Hensher and Swait [2000; p.38]), and in the stochastic choice literature they also refer to trembles or errors by the individual (e.g., Hey [1995]).

<sup>37</sup> This is particularly clear in the exposition of Louviere, Hensher and Swait [2000; p. 237, 244] since they use the notation  $\alpha^{RP}$  and  $\alpha^{SP}$  for the intercepts from data sources RP and SP, and a common  $\beta$  for the pooled estimates.

the real data. Since the data are pooled, the same implicit normalization of variance is applied to the two data sources. Thus one effectively constrains the variance normalizations to be the same, but allows the intercept to vary according to the data source. The hypothesis of interest is then tested by means of an appropriate comparisons of likelihood values.

In effect, this procedure can test if hypothetical and real responses are affected by covariates in the same manner, but not if they differ conditional on the covariates. Thus if men and women have the same propensity to purchase a good at some price, this method can identify that. But if men and women each have the same elevated propensity to “purchase” when the task is hypothetical, this method will not identify that.<sup>38</sup> And the overall likelihood tests will indicate that the data can be pooled, since the method allows the intercepts to differ across the two data sources. Hence claims in Louviere, Hensher and Swait [2000; ch.13] of widespread “preference regularity” across disparate data sources and elicitation methods should be read with a grain of salt in terms of the implications for the need to calibrate hypothetical and real responses.<sup>39</sup>

On the other hand, the *tests* of preference regularity from the marketing literature are capable of being applied more generally than the methods of *pooling* preferences from different sources. The specifications considered by Louviere, Hensher and Swait [2000; p. 233-236] clearly admit the possibility of marginal valuations differing across hypothetical and real settings.<sup>40</sup> In fact, it is possible to undertake tests that some coefficients are the same while others are different, illustrated by Louviere, Hensher and Swait [2000; §8.4.2]. This is a clear analogue to some parameters in a real/hypothetical experiment being similar (e.g. some marginal effects) but others being quite different (e.g. purchase intention), as illustrated by Lusk and Schroeder [2004]. The appropriate pooling procedures then allow some coefficients to be estimated jointly while others are estimated separately, although there is an obvious concern with such specification

---

<sup>38</sup> Interactions may or may not be identified, but they only complicate the already-complicated picture.

<sup>39</sup> Despite this negative assessment of the potential of this approach for constructive calibration of differences between hypothetical and real responses, the “data enrichment” metaphor that originally motivated this work in marketing is an important and fundamental one for economics.

<sup>40</sup> Louviere, Hensher and Swait [2000; p. 233] use the notation  $\alpha^{RP}$  and  $\alpha^{SP}$  for the intercepts from data sources RP and SP, and  $\beta^{RP}$  and  $\beta^{SP}$  for the coefficient estimates.

tests leading to reported standard errors that understate the uncertainty over model specification.

### 3.3 Calibrating Responses Within-Sample

Fox et al. [1994] and List and Shogren [1998][2002] propose a method of calibration which uses hypothetical and real responses from the same subjects for the *same good*.<sup>41</sup> But if one is able to elicit values in a non-hypothetical manner, then why bother in the first place eliciting hypothetical responses that one has to calibrate? The answer is that the relative cost of collecting data may be very different in some settings. It is possible in marketing settings to construct a limited number of “mock ups” of the potential product to be taken to market, but these are often expensive to build due to the lack of scale economies. Similarly, one could imagine in the environmental policy setting that one could actually implement policies on a small scale at some reasonable expense, but that it is prohibitive to do so more widely without some sense of aggregate WTP for the wider project. The local implementation could then be used as the basis for ascertaining how one must adjust hypothetical responses for the wider implementation.

These considerations aside, the remaining substantive challenge for calibration is to demonstrate feasibility and utility for the situation of most interest in environmental valuation, when the underlying target good or project is non-deliverable and one must by definition consider cross-commodity calibration.

## 4. Conclusions

### 4.1 The Problem of Hypothetical Bias

Hypothetical bias is a problem that must be dealt with squarely.<sup>42</sup> It is arguable that many of the “ad hoc” practices of the survey literature, and attendant “fixes” for certain problems, derive from a

---

<sup>41</sup> Fox et al. [1994; p.456] offer two criticisms of the earlier calibration approach of Blackburn et al. [1994]. The first is that it is “inconclusive” since one of the bias functions has relatively large standard errors. But such information on the imprecision of valuations is just as important as information on the point estimates if it correctly conveys the uncertainty of the elicitation process. In other words, it is informative to convey one’s imprecision in value estimation if the decision-maker is not neutral to risk. The second criticism is that Blackburn et al. [1994] only elicit a calibration function for one price on a demand schedule in their illustration of their method, and that the calibration function might differ for different prices. This is certainly correct, but hardly a fundamental criticism of the method in general.

<sup>42</sup> There are some settings in which the issue of hypothetical bias rages on, but is simply a red herring. A prime example is the continuing debate over the role of hypothetical responses in tests of expected utility theory.

systematic avoidance of looking into the bright light of hypothetical bias. With the controlled sunglasses of experimental methods, however, it is possible to tackle this problem and avoid the embarrassed shuffling that surrounds it in the extant survey literature.

Consider, as an important example, the “WTA *versus* WTP” controversy. One of the most embarrassing gaps in the state-of-the-art of environmental damage assessment is the avoidance of WTA measures. It is a gap because WTA is more often than not the natural legal counterpart to the issue under study, rather than a WTP measure. The use of “unnatural” counterparts leads to difficulties in the design of credible and correct scenarios for value elicitation: “why should I pay to avoid another *Exxon Valdez* oil spill when Exxon should be paying *me* for the last one?!” This problem is heightened by the unfortunate tendency to delete the more articulate and thoughtful of respondents, reacting to such illogic, as “protest bids.”

The gap in the literature and practice of environmental damage assessment is intellectually embarrassing, given that there is no coherent rationale for avoiding WTA. The reality is that readers of reports using the CVM are being good Bayesians: the WTA numbers reported on hypothetical surveys conflict *so* violently with our priors, as to what real economic commitment individuals would actually make, that we simply choose to ignore them. Our priors outweigh the data, and justifiably so.

But why not apply the same priors to the WTP numbers and throw them out as well? One could argue that the only problem with WTA is that it exhibits a substantial hypothetical bias. However, the literature has demonstrated that WTP suffers from a substantial hypothetical bias as well. Given that hypothetical bias is also a problem for WTP measures, and must be corrected for if *either* measure is to have any credibility, why has the CVM literature simply chosen to avoid WTA? It is not the case that WTP values elicited from hypothetical surveys are reliable while hypothetical WTA values are not: neither set of values is unambiguously reliable as they stand now.

It may turn out that the hypothetical bias of WTA is larger than the hypothetical bias of WTP, if the conventional wisdom is correct. However, this says nothing about whether or not the bias in WTP is easier to correct for. In other words, the bias in WTA may be larger in size but it could also be more

predictable. As argued by Blackburn, Harrison and Rutström [1994], it is the (statistical) predictability of the bias that is crucial to our ability to correct for it.

The general point which this example illustrates is that many of the priors of CVM researchers have become so ingrained into the “state of the art” that they are rarely questioned. Even worse than not confronting those practices, one often sees a hardening of the intellectual arteries when rationalizing them.<sup>43</sup>

My judgement is that we have a relatively good handle on the problem of *hypothetical bias*. This does not mean that we know how to reduce it to zero, but we do know how to mitigate it by adroit and tested choice of words and how to adjust our results for it when present. Given the vested interests of industry groups in having lower environmental damage estimates gain some popularity and credibility, I see no likely shortage of interest in this line of research. Apart from the exotic prospects of lab experimenters being well-paid to ply their craft in the field, there is nothing of great methodological interest remaining to be done in this arena after the first series of field applications of the basic lab ideas and designs have been undertaken.

## 4.2 Best-Practice Experimental Methods

What practices emerge as best, where “best” is measured with respect to the ability to claim control? The only contribution of experimental methods is the ability to stake some *a priori* claim to having controlled certain factors that might otherwise contaminate non-experimental data. Those factors *might* not be a problem, but in the absence of controls for their absence or mitigation we simply do not know and must rely on assertion. To paraphrase Smith [1982; p.938] it is precisely this type of frustration with the

---

<sup>43</sup> For example, Carson [1991; p.130] correctly notes that “... the debate on WTP versus WTA is sometimes decided in favor of WTP based on questionable logic, such as the following: WTA is the correct measure but since it cannot be measured, the researcher should measure WTP instead. This logic was adopted, for instance, in the U.S. Department of Interior [1986] natural resource damage assessment guidelines ....” However, one then finds essentially the same questionable logic employed in the *Exxon Valdez* case by Carson et al. [1992; p.7]: “Thus, willingness to accept compensation is the theoretically correct measure in this case. Unfortunately, it is very difficult to design a survey that effectively elicits WTA amounts because respondents tend to regard WTA scenarios as implausible. Therefore, in the current damage assessment, we chose willingness to pay as the valuation framework even though this choice will understate the true value of losses suffered as a result of the spill...” (footnote deleted).

need to rely on assertion which prompted so many researchers to turn to using experimental methods to evaluate the CVM. Moreover, these problems of control are likely to become even more severe as one ventures out of the laboratory and into the field, as experimenters are doing now.<sup>44</sup>

First, one should use one-shot institutions rather than repeated institutions with the same good. Specifically, if a Vickrey auction is to be employed it is better to use a one-shot Vickrey auction than to have many repetitions of a Vickrey auction with the same good.<sup>45</sup> There is some advantage in repetitions of a Vickrey auction in terms of training subjects in the logistical procedures of the experiment or the specific dominant strategy properties of the institution being used. In these respects it is possible to train subjects up in another good altogether such as the tasty choice of (Hershey's) Kisses by Coller and Williams [1999]. Similarly, it is possible to simply inform subjects of the dominant strategy property of the institution they are using, such as in Neill et al. [1994] and Rutström [1998].

Can one just do repeated Vickrey auctions and report the results of the first period for those skeptics, like me, who dismiss the later periods as potentially contaminated? Unfortunately not. The problem is that the subjects typically are told that the experiment will last for several periods, and we do not then know how they decide to bid in the first round. It is perfectly plausible that a subject might understand the logic of a dominant strategy for a one-shot auction but not be able to see that this logic applies equally to each of the stage games of a repeated (experimental) game.

The second implication for best practice experimental design is the importance of having simultaneous bid submission rather than having real-time bid submission or real-time sequential bid submission such as one would find in an English auction. The implication here is that one might best use

---

<sup>44</sup> Harrison and List [2004] discuss the ways in which field experiments differ from laboratory experiments.

<sup>45</sup> Loomes, Starmer and Sugden [2003] correctly note that affiliation in values may be a confound in some recent experimental studies claiming that anomalies disappear in “real markets” that have stationary repetition. Their own efforts to design a procedure that allows repetition but avoids what they call “the shaping hypothesis” illustrate some of the difficulties that can be best avoided by using one-shot elicitation procedures. Moreover, their shaping hypothesis is only one possible way in which past prices can influence current valuations. Harrison, Harstad and Rutström [2004] provide a more general set of hypotheses about these effects. The danger is that even if one can design around one particular hypothesis of affiliation, such as their shaping hypothesis, others may still be at work to confound inferences.

sealed-bid institutions rather than standard forms of sequential-bid, or real-time, bid institutions.<sup>46</sup> In the context of revealed preference experiments, in which multiple choices are presented to the subject, space constraints prohibit the simultaneous presentation of all choices. In this case, randomization of order should serve to control for any order effects.

A third implication for experimental design is to try to build in some controls for field substitutes. In this respect it is important to distinguish between the subjective transactions costs that individuals might have in going from the lab to the field and the subjective beliefs that individuals might have with respect to the price of field substitutes. Each of these are important dimensions of the problem of controlling for field substitutes, and each is to some extent amenable to experimental control. Attempts to do that by way of eliciting information on subject characteristics are provided by Coller and Williams [1999] and Harrison, Lau and Williams [2002].

The fourth implication is the importance of collecting information routinely on a standard array of individual subject characteristics. Although student samples tend to have statistically degenerate characteristics, it is a relatively easy matter to recruit samples outside colleges.

A fifth implication is that designs should be developed that allow subjects to experience different institutional variants, with as many of them as possible being from the class of incentive compatible institutions. The reason is that there are likely to be differences in the behavioral responses that different subjects have in different institutions: some people understand the logic of a one-shot Vickrey auction immediately, others never do; some people understand numbers, others need graphs. It is quite likely that one can identify these subject types, and either (a) select formally equivalent institutions that will elicit more reliable responses for different subjects, or (b) correct for expected biases in responses. In short, avoid looking for a “magic bullet” institution that will work for every individual in every context.

More generally there are two implications for experimental design that are worth stating even if

---

<sup>46</sup> Non-standard institutions could be easily devised, in which subjects could revise their bids in real-time and yet not be informed of the prior bids of other agents. One mitigating factor in favor of real-time English auctions is that subjects appear to understand their dominant strategy better in that setting than in one-shot sealed-bid auctions, as demonstrated by Rutström [1998] and Harstad [2000].

they sound a little more abstract. The first is the great danger of engaging in mixing and matching of different institutions in experimental economics. This problems bedevils earlier attempts at value elicitation, particularly those pertaining to environmental damage assessment and the CVM. Many of these experiments used some of the procedures of institutions that are demand revealing and some of the procedures of other institutions that are demand revealing, where the two sets of institutions were quite distinct. Thus one obtained the flavor of a witches' brew: add a dab of Vickrey, add a pinch of Smith auction, add a liberal dose of repetition to ensure researching of the preferences, and stir liberally. Such concoctions are destined to provide game theorists with nightmares. Although this may be a worthwhile externality, it does not make for robust science. The subtleties discussed above should provide ample evidence as to the dangers of casual application of rigorous notions of incentive compatibility.<sup>47</sup>

The final implication is an important one for those researchers in the environmental valuation literature that are discovering experimental methods for the first time. That implication is simply to read the extant literature critically. Experimental economics has progressed dramatically in the last two decades, most notably by successfully integrating careful economic theory with the goals of applied microeconomics. Many mistakes have been made, and many remain to be made. It is disconcerting, however, to observe the propensity for known mistakes to be actively propagated. Relative to some of the serious problems which confront the field of environmental damage assessment, untangling the knotted threads of extant experimental practice is trivial.

---

<sup>47</sup> The worst examples of these design problems are discussed in Cummings and Harrison [1994].

## References

- Aadland, David, and Caplan, Arthur J., "Willingness to Pay for Curbside Recycling with Detection and Mitigation of Hypothetical Bias," *American Journal of Agricultural Economics*, 85, 2003, 492-502.
- Afriat, Sidney, "The Construction of a Utility Function from Expenditure Data," *International Economic Review*, 8, 1967, 67-77.
- Arrow, Kenneth; Solow, Robert; Portney, Paul; Leamer, Edward E.; Radner, Roy; and Schuman, Howard, "Report of the NOAA Panel on Contingent Valuation," *Federal Register*, 58(10), January 15, 1993, 4602-4614.
- Blackburn, McKinley; Harrison, Glenn W., and Rutström, E. Elisabet, "Statistical Bias Functions and Informative Hypothetical Surveys," *American Journal of Agricultural Economics*, 76(5), December 1994, 1084-1088.
- Blumenschein, Karen; Johannesson, Magnus; Blomquist, Glenn C.; Liljas, Bengt, and O'Coner, Richard M., "Experimental Results on Expressed Certainty and Hypothetical Bias in Contingent Valuation," *Southern Economic Journal*, 65, 1998, 169-177.
- Blumenschein, Karen; Johannesson, Magnus; Yokoyama, Krista K, and Freeman, Patricia R., "Hypothetical Versus Real Willingness to Pay in the Health Care Sector: Results from a Field Experiment," *Journal of Health Economics*, 20, 2001, 441-457.
- Bohm, Peter; Lindén, Johan, and Sonnegård, Joakim, "Eliciting Reservation Prices: Becker-DeGroot-Marschak Mechanisms vs. Markets," *Economic Journal*, 107, July 1997, 1079-1089.
- Boyce, Rebecca R.; Brown, Thomas C.; McClelland, Gary H.; Peterson, George; and Schulze, William D., "Experimental Evidence of Existence Value in Payment and Compensation Contexts," in Kevin J. Boyle, and T. Heekin (eds.), *Western Regional Research Project W-133: Benefits and Costs in Natural Resources Planning*, Interim Report 2, Department of Agricultural and Resource Economics, University of Maine, Orono, July 1, 1989.
- Brown, Thomas C.; Ajzen, Icek, and Hrubes, Daniel, "Further Tests of Entreaties to Avoid Hypothetical Bias in Referendum Contingent Valuation," *Journal of Environmental Economics and Management*, 46(2), September 2003, 353-361.
- Cameron, Trudy Ann; Poe, Gregory L.; Ethier, Robert G., and Schulze, William D., "Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same?" *Journal of Environmental Economics and Management*, 44, 2002, 391-425.
- Carlsson, Fredrick, and Martinsson, Peter, "Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments?" *Journal of Environmental Economics and Management*, 41, 2001, 179-192.
- Carson, Richard T., "Constructed Markets," in J.B. Braden and C.K. Kolstad (eds.), *Measuring the Demand for Environmental Quality* (Amsterdam: North-Holland, 1991).
- Carson, Richard T; Hanemann, W. Michael; Krosnick, Jon A.; Mitchell, Robert C.; Presser, Stanley; Ruud, Paul A., Smith, V. Kerry, "Referendum Design and Contingent Valuation: The NOAA Panel's No-Vote Recommendation," *Review of Economics and Statistics*, 80(2), May 1998, 335-338; reprinted with typographical corrections in *Review of Economics and Statistics*, 80(3), August 1998.

- Carson, Richard T.; Mitchell, Robert C.; Hanemann, W. Michael; Kopp, Raymond J.; Presser, Stanley; and Ruud, Paul A., *A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill* (Anchorage: Attorney General of the State of Alaska, November 1992).
- Coller, Maribeth, and Williams, Melonie B., "Eliciting Individual Discount Rates," *Experimental Economics*, 2, 1999, 107-127.
- Cummings, Ronald G.; Elliott, Steven; Harrison, Glenn W., and Murphy, James, "Are Hypothetical Referenda Incentive Compatible?" *Journal of Political Economy*, 105(3), June 1997, 609-621.
- Cummings, Ronald G., and Harrison, Glenn W., "Was the *Ohio* Court Well Informed in Their Assessment of the Accuracy of the Contingent Valuation Method?" *Natural Resources Journal*, 34(1), Winter 1994, 1-36.
- Cummings, Ronald G.; Harrison, Glenn W., and Osborne, Laura L., "Can the Bias of Contingent Valuation Be Reduced? Evidence from the Laboratory," *Economics Working Paper B-95-03*, Division of Research, College of Business Administration, University of South Carolina, 1995a (see <http://www.bus.ucf.edu/gharrison/wp/>).
- Cummings, Ronald G.; Harrison, Glenn W., and Osborne, Laura L., "Are Realistic Referenda Real?" *Economics Working Paper B-95-06*, Division of Research, College of Business Administration, University of South Carolina, 1995b (see <http://www.bus.ucf.edu/gharrison/wp/>).
- Cummings, Ronald G.; Harrison, Glenn W., and Rutström, E. Elisabet, "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive Compatible?" *American Economic Review*, 85(1), March 1995, 260-266.
- Cummings, Ronald G. and Taylor, Laura O., "Does Realism Matter in Contingent Valuation Surveys?" *Land Economics*, 74(2), 1998, 203-215.
- Cummings, Ronald G. and Taylor, Laura O., "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method," *American Economic Review*, 89(3), June 1999, 649-665.
- Dillman, Don A., *Mail and Telephone Surveys: The Total Design Method* (New York: Wiley, 1978).
- Fischhoff, Baruch, "Value Elicitation. Is there Anything in There?," *American Psychologist*, 46, August 1991, 835-847.
- Fischhoff, Baruch, and Furby, L., "Measuring Values: A Conceptual Framework for Interpreting Transactions with Special Reference to Contingent Valuations of Visibility," *Journal of Risk and Uncertainty*, 1, 1988, 147-184.
- Fox, John A.; Shogren, Jason F.; Hayes, Dermot J., and Kliebenstein, James B., "CVM-X: Calibrating Contingent Values with Experimental Auction Markets," *American Journal of Agricultural Economics*, 80, August 1998, 455-465.
- Freeman III, A. Myrick, *The Benefits of Environmental Improvement* (Baltimore: Johns Hopkins Press, 1979).
- Haab, Timothy C.; Huang, Ju-Chin, and Whitehead, John C., "Are Hypothetical Referenda Incentive Compatible? A Comment," *Journal of Political Economy*, 107(1), February 1999, 186-196.
- Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions: Reply," *American Economic Review*, 82, December 1992, 1426-1443.

- Harrison, Glenn W.; Beekman, Robert L.; Brown, Lloyd B.; Clements, Leianne A.; McDaniel, Tanga M.; Odom, Sherry L. and Williams, Melonie, "Environmental Damage Assessment With Hypothetical Surveys: The Calibration Approach," in M. Boman, R. Brännlund and B. Kriström (eds.), *Topics in Environmental Economics* (Amsterdam: Kluwer Academic Press, 1999).
- Harrison, Glenn W., Harstad, Ronald M., and Rutström, E. Elisabet, "Experimental Methods and Elicitation of Values," *Experimental Economics*, 7(2), June 2004, 123-140.
- Harrison, Glenn W.; Lau, Morten Igel, and Williams, Melonie B., "Estimating Individual Discount Rates for Denmark: A Field Experiment," *American Economic Review*, 92(5), December 2002, 1606-1617.
- Harrison, Glenn W.; Lau, Morten Igel; Rutström, E. Elisabet, and Sullivan, Melonie B., "Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues," in J. Carpenter, G.W. Harrison and J.A. List (eds.), *Field Experiments in Economics* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 10, 2005).
- Harrison, Glenn W., and Lesley, James C., "Must Contingent Valuation Surveys Cost So Much?" *Journal of Environmental Economics and Management*, 31, June 1996, 79-95.
- Harrison, Glenn W., and List, John A., "Field Experiments," *Journal of Economic Literature*, 42(4), December 2004, 1013-1059.
- Harrison, Glenn W., and Rutström, E. Elisabet, "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods," in C.R. Plott and V.L. Smith (eds.), *Handbook of Experimental Economics Results*, North-Holland: Amsterdam, 2005.
- Harstad, Ronald M., "Dominant Strategy Adoption and Bidders' Experience with Pricing Rules," *Experimental Economics*, 3(3), December 2000, 261-280.
- Hayes, Dermot J.; Shogren, Jason; Shin, Seung Y., and Kliebenstein, James B., "Valuing Food Safety in Experimental Auction Markets," *American Journal of Agricultural Economics*, 77, February 1995, 40-53.
- Hensher, David; Louviere, Jordan, and Swait, Joffre D., "Combining Sources of Preference Data," *Journal of Econometrics*, 89, 1999, 197-221.
- Hey, John D., "Experimental Investigations of Errors in Decision Making Under Risk," *European Economic Review*, 39, 1995, 633-640.
- Hoffman, Elizabeth; Menkhous, Dale J.; Chakravarti, Dipinkar; Field, Ray A., and Whipple, Glen D., "Using Laboratory Experimental Auctions in Marketing Research: A Case Study of New Packaging for Fresh Beef," *Marketing Science*, 12(3), Summer 1993, 318-338.
- Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), December 2002, 1644-1655.
- Horowitz, John K., "Discounting Money Payoffs: An Experimental Analysis," *Handbook of Behavioral Economics* (Greenwich, CT: JAI Press, Inc., v. 2B, 1991, 309-324).
- Imber, David; Stevenson, Gay; and Wilks, Leanne, *A Contingent Valuation Survey of the Kakadu Conservation Zone* (Canberra: Australian Government Publishing Service for the Resource Assessment Commission, February 1991).
- Johannesson, Magnus; Blomquist, Glenn C.; Blumenschein, Karen; Johansson, Per-Olov; Liljas, Bengt, and O'Conner, Richard M., "Calibrating Hypothetical Willingness to Pay Responses," *Journal of*

- Risk and Uncertainty*, 8, 1999, 21-32.
- Kahneman, Daniel; Knetsch, Jack L., and Thaler, Richard H., "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy*, 98, December 1990, 1325-1348.
- Kagel, John H.; Harstad, Ronald M., and Levin, Dan, "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study," *Econometrica*, 55, November 1987, 1275-1304.
- Krosnick, Jon A.; Holbrook, Allyson L.; Berent, Matthew K.; Carson, Richard T.; Hanemann, W. Michael; Kopp, Raymond J.; Mitchell, Robert C.; Presser, Stanley; Ruud, Paul A. Ruud; Smith, V. Kerry; Moody, Wendy R.; Green, Melanie C., and Conaway, Michael, "The Impact of 'No Opinion' Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice?" *Public Opinion Quarterly*, 66, 2002, 371-403.
- Kurz, Mordecai, "Experimental Approach to the Determination of the Demand for Public Goods," *Journal of Public Economics*, 3, 1974, 329-348.
- List, John A., "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *American Economic Review*, 91(5), December 2001, 1498-1507.
- List, John A.; Berrens, Robert P.; Bohara, Alok K., and Kerkvliet, Joe, "Examining the Role of Social Isolation on Stated Preferences," *American Economic Review*, 94(3), June 2004, 741-752.
- List, John A., and Shogren, Jason F., "Calibration of the Differences Between Actual and Hypothetical Valuations in a Field Experiment," *Journal of Economic Behavior and Organization*, 37, November 1998, 193-205.
- List, John A., and Shogren, Jason F., "Price Signals and Bidding Behavior in Second-Price Auctions with Repeated Trials," *American Journal of Agricultural Economics*, 81, November 1999, 942-929..
- List, John A., and Shogren, Jason F., "Calibration of Willingness-to-Accept," *Journal of Environmental Economics and Management*, 43(2), 2002, 219-233.
- Loomes, Graham; Starmer, Chris, and Sugden, Robert, "Do Anomalies Disappear in Repeated Markets?," *Economic Journal*, 113, March 2003, C153-C166.
- Loomis, John; Brown, Thomas; Lucero, Beatrice, and Peterson, George, "Improving Validity Experiments of Contingent Valuation Methods: Results of Efforts to Reduce the Disparity of Hypothetical and Actual Willingness to Pay," *Land Economics*, 72(4), November 1996, 450-461.
- Loomis, John; Gonzalez-Caban, Armando, and Gregory, Robin, "Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates?" *Land Economics*, 70, November 1994, 499-506.
- Louviere, Jordan J.; Hensher, David A., and Swait, Joffre D., *Stated Choice Methods: Analysis and Application* (New York: Cambridge University Press, 2000).
- Lusk, Jayson L., and Schroeder, Ted C., "Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks," *American Journal of Agricultural Economics*, 86(2), May 2004, 467-482.
- McClelland, Gary; Schulze, William; Waldman, Donald; Irwin, Julie; and Schenk, David, "Sources of Error in Contingent Valuation," *Unpublished Manuscript*, Department of Economics, University of

- Colorado at Boulder, January 1991.
- Milgrom, Paul R., and Weber, Robert J. "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(5), September 1982, 1089-1122.
- Mitchell, Robert C., and Carson, Richard T., *Using Surveys to Value Public Goods: The Contingent Valuation Method* (Baltimore: Johns Hopkins Press, 1989).
- Murphy, James J.; Stevens, Thomas, and Weatherhead, Darryl, "An Empirical Study of Hypothetical Bias in Voluntary Contribution Contingent Valuation: Does Cheap Talk Matter?" *Unpublished Manuscript*, Department of Resource Economics, University of Massachusetts, July 2003.
- Nape, Steven W.; Frykblom, Peter; Harrison, Glenn W., and Lesley, James C., "Hypothetical Bias and Willingness to Accept," *Economic Letters*, 78(3), March 2003, 423-430.
- National Oceanic and Atmospheric Administration, "Contingent Valuation Panel, Public Meeting, Wednesday, August 12, 1992," *Certified Official Transcript*, 283 pp. plus attachments, Department of Commerce, Washington, DC., 1992.
- National Oceanographic and Atmospheric Administration, "Proposed Rules for Valuing Environmental Damages," *Federal Register*, 59(5), January 7, 1994a, 1062-1191.
- National Oceanic and Atmospheric Administration, "Natural Resource Damage Assessments; Proposed Rules," *Federal Register*, 59(85), May 4, 1994b (Part II).
- Neill, Helen R.; Cummings, Ronald G.; Ganderton, Philip T.; Harrison, Glenn W., and McGuckin, Thomas, "Hypothetical Surveys, Provision Rules and Real Economic Commitments," *Land Economics*, 70(2), May 1994, 145-154.
- Rowe, Robert D.; Schulze, William D.; Shaw, W. Douglass; Schenk, David, and Chestnut, Lauraine G., "Contingent Valuation of Natural Resource Damage Due to the Nestucca Oil Spill," *Final Report*, RCG/Hagler, Bailly, Inc., Boulder, June 15, 1991.
- Rutström, E. Elisabet, "Home-Grown Values and the Design of Incentive Compatible Auctions," *International Journal of Game Theory*, 27(3), 1998, 427-441.
- Samuelson, Paul A., "A Note on the Pure Theory of Consumer's Behavior," *Economica*, 5(17), February 1938, 61-71.
- Schkade, David A., and Payne, John W., "Where Do the Numbers Come From? How People Respond to Contingent Valuation Questions," in J. Hausman (ed.), *Contingent Valuation: A Critical Appraisal* (Amsterdam: North-Holland, 1993).
- Shogren, Jason F., "Experimental Methods and Valuation," in K-G. Mäler and J. Vincent (eds.), *Handbook of Environmental Economics. Volume 2: Valuing Environmental Changes* (Amsterdam: North-Holland, 2004).
- Shogren, Jason F.; Shin, Seung Y.; Hayes, Dermot J., and Kliebenstein, James B., "Resolving Differences in Willingness to Pay and Willingness to Accept," *American Economic Review*, 84(1), March 1994, 255-270.
- Shogren, Jason F.; List, John A., and Hayes, Dermot J., "Preference Learning in Consecutive Experimental Auctions," *American Journal of Agricultural Economics*, 82(4), November 2000, 1016-1021.

- Smith, V. Kerry, "Can We Measure the Economic Value of Environmental Amenities?," *Southern Economic Journal*, 56, 1990, 865-887.
- Smith, V. Kerry, "Arbitrary Values, Good Causes, and Premature Verdicts," *Journal of Environmental Economics and Management*, 22, January 1992, 71-89.
- Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, 72(5), December 1982, 923-955.
- Svedsäter, Henrik, and Johansson-Stenman, Olof, "Choice experiments and Self Image: Hypothetical and Actual Willingness To Pay," *Unpublished Manuscript*, Department of Economics, Gothenburg University, 2001.
- Swait, Joffre, and Louviere, Jordan, "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research*, 30, August 1993, 305-314.
- Varian, Hal R., "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, July 1982, 945-73.
- Varian, Hal R., "Non-Parametric Tests of Consumer Behavior," *Review of Economic Studies*, 50, January 1983, 99-110.