Chapter 81

EXPERIMENTAL EVIDENCE ON THE EXISTENCE OF HYPOTHETICAL BIAS IN VALUE ELICITATION METHODS

GLENN W. HARRISON and E. ELISABET RUTSTRÖM

Hypothetical bias is said to exist when values that are elicited in a hypothetical context, such as a survey, differ from those elicited in a real context, such as a market. Although reported experimental results are mixed, we claim that the evidence strongly favor the conclusion that hypothetical bias exists. Nevertheless, some optimism is called for with respect to the ability to calibrate either the hypothetical survey instrument itself, or the responses from it. Calibration could be performed in such a way that elicited values are unbiased estimates of actual valuations.

The usual concern with hypothetical bias is that people will overstate their true valuation in hypothetical settings. Why? As a matter of logic, if you do not have to pay for the good but a higher *verbal* willingness to pay (WTP) response increases the chance of it's provision, then verbalize away to increase your expected utility! Similarly, if your verbal statement of a higher willingness to accept (WTA) compensation might reduce the risk that the commodity will be taken away from you, verbalize as high an amount as you want. This seems like such a simple logic, so why even bother testing if the problem arises behaviorally with experiments? The rationale for these tests arises from debates over the validity of the Contingent Valuation Method (CVM) for valuing environmental damages. The CVM is a survey instrument commonly used in environmental resource valuations, which is hypothetical in both the payment for and the provision of the good being valued. Proponents of the CVM have often claimed that hypothetical bias simply does not exist or that it is quantitatively unimportant even if it does exist.[1] Given the importance that such claims about the validity of the CVM have for environmental policy, and the confidence that some proponents of the CVM place in their *a priori* claims, testing the existence of hypothetical bias has been necessary. Moreover, tests that establish the existence of the behavioral problem invariably provide information to help one quantify how important it is. That can also be useful when measuring the efficacy of attempts to mitigate the problem.

Some simple experimental tests have managed to shift the burden of proof back on those making claims that hypothetical bias is not a problem. In addition, several experimental results point the way to a more complete understanding of the process of eliciting values, as well as constructive means of mitigating the effects of hypothetical bias. The usual scenario in experiments has been the exchange of a commodity, although for policy purposes the scenario is often some change in policy. The use of a commodity

[1] See Cummings and Harrison (1994) for a literature review of the CVM debates over hypothetical bias.

Handbook of Experimental Economics Results, Volume 1

DOI: 10.1016/S1574-0722(07)00081-9

instead of some general policy change simplifies the problem of detecting and dealing with hypothetical bias. The reason is that the description of the scenario itself may entail hypothetical or incredible features (e.g., "are you willing to pay $5 to eradicate the risk of global warming"). In this case one might easily confound the perception of a non-credible policy with the hypothetical nature of the payment itself. Thus the use of simple commodities in experiments should be seen as an instance of the experimenter gaining more control over the behavior under study.

The experimental literature on the existence of hypothetical bias falls into three groups. The first group of studies examined open ended elicitation methods in which the object to be valued is presented to subjects and a dollar value solicited. The typical institution used in these studies is the sealed-bid, second-price (Vickrey) auction. The second group of studies examined closed ended elicitation methods (often called dichotomous choice) in which the object to be valued and an offer price is presented to subjects, and a purchase decision solicited. The typical institution used in these studies is the posted offer institution. The third group of studies examined binary choice in a referendum context. A group of subjects is offered a public good at a given price and asked to vote yes or no.

In all cases the results are quite clear: there is a hypothetical bias problem to be solved. We review the basic experimental results that support this conclusion. Table 1 summarizes these results. The conclusion is that the weight of the evidence supports the claim that hypothetical valuations exceed real valuations.

The literature of experimental results falls into two waves. The first consists of a series of relatively disconnected studies in the 1980's using innovative experimental designs in an attempt to shed light on the general validity of the CVM. These are briefly reviewed in Sections 1 and 2. More recently there has been a series of experimental studies using extremely simple designs that are more narrowly focused on the issue of hypothetical bias and how it interacts with the elicitation mechanism. These are reviewed in Sections 3 through 4.

1. The CVM Literature and Tests with Private Goods

We begin with a study that claims not to have found any significant hypothetical bias, although the difference in responses between hypothetical and actual contexts is large. Dickie, Fisher, and Gerking (1987) obtained values for a pint of strawberries by using CVM, and also by actually selling the strawberries to households. They concluded (p. 73) that they could not reject the null hypothesis of structurally identical demand equations estimated from actual sales and CVM data. However, their results are mixed if one examines them in further detail. Specifically, they found that there were large differences in the estimated actual and hypothetical demand schedules (p. 74, in their Tables 3 and 4, and Figure 1). Depending on the price one uses, the hypothetical demand curve can overstate the quantity demanded from 35 1.4% to 68.6% if all interview team results are included. The average hypothetical bias, calculated from the raw responses,

Table 1
Summary of findings of hypothetical bias

Study	Key features of bias comparison	Hypothetical bias	Statistically significant?
BH		−46%	no
JLJ	Conservative interpretation of CVM responses vs real; largest bias reported.	−43%	no
MSC	Smallest bias found at highest risk treatment.	−25%	?
Bohm	Comparison with auction institution.	0%	no
JLJ	Conservative interpretation of CVM responses vs real; smallest bias reported.	0%	no
SM		3%	no
DFG	Excluding team 2 and outlier identified in study.	8%	no
Bohm	Comparison of means; smallest bias reported.	16%	yes
JLJ	Standard interpretation of CVM responses vs real; smallest bias for lowest price.	19%	yes
Bohm	Comparison of medians; smallest bias reported.	25%	?
Griffin et al.	Comparison of proportion of "yes" responses.	29%	?
KMD	Private good.	30%	yes
DP	Means, excluding non-respondents.	35%	?
Bohm	Comparison of means; highest bias reported.	40%	yes
Frykblom	Dichotomous choice responses.	56%	yes
DFG	Comparison of "raw" means.	58%	no
BHK	Comparison of means.	60%	yes
Frykblom	Open-ended responses.	60%	yes
CEHM		67%	yes
DFG	Estimated demand functions; comparison of means; smallest bias reported.	69%	no
KMP	Public good.	100%	yes
Bohm	Comparison of medians; largest bias reported.	100%	?
MSC	Largest bias reported at lowest risk.	120%	?
CHR	Juicers; within-subjects.	163%	yes
CHR	Calculators; between-subjects.	163%	yes
BHK	Comparison of medians.	176%	?
DP	Comparison of means, assuming non-respondents have WTP = $0.	203%	?
Neill et al.	Painting; comparison of means.	290%	yes
DFG	Estimated demand functions; comparison of means; largest bias reported.	351%	no
Neill et al.	Painting; comparison of medians.	400%	?
JLJ	Standard interpretation of CVM responses vs real; highest bias reported.	701%	yes
CHR	Chocolates; within-subjects comparisons.	873%	yes
SS	Comparison of means.	2017%	?

(*continued on next page*)

Table 1
(*continued*)

Study	Key features of bias comparison	Hypothetical bias	Statistically significant?
Neill et al.	Map; HVA vs RVA; comparison of means.	2400%	yes
Neill et al.	Map; CVM vs RVA; comparison of means.	2600%	yes

Notes: The acronym for each study is provided in the References; bias measured as hypothetical minus real over real when positive, and real minus hypothetical over hypothetical when negative; statistical significance is determined at a one-sided 10% level, and a question mark indicates that there was insufficient information in the study to allow a simple determination.

is about 58%.[2] We include this number in Table 1 that compiles the results from all the studies we survey. We conclude that there is hardly unequivocal support in this study for the view that hypothetical and actual questions generate the same demand schedules.

Furthermore, a number of other studies demonstrate that there may be significant differences between CVM values and market values.

Bishop and Heberlein (1979) and Bishop, Heberlein, and Kealy (1983) found significant differences between CVM estimates for subjects' WTA for returning their goose hunting permits and WTA values based upon actual cash payments. The expected value based on actual cash payments was \$63 and that based on the CVM was \$101, implying a bias of about 60%, as shown in Table 1. Median values were \$29 and \$80 for the actual and hypothetical valuations, respectively, a bias of 176%. Mitchell and Carson (1989, pp. 195–199) dispute these conclusions. Hanemann (1984) also re-evaluates the results of Bishop and Heberlein, demonstrating the extreme sensitivity of their conclusions to alternative statistical assumptions.

In a later study, Bishop and Heberlein (1986) obtained CVM and actual cash values for deer hunting permits, and found that WTP values were significantly overstated in the CVM relative to the cash market. In three different valuation institutions they found that the average hypothetical values exceeded the real ones by 33%, 68%, and 79%. In one institution did they find that the real valuations exceeded the hypothetical by 46%, but this was not significant.

2. The CVM Literature and Tests with Public Goods

We next turn to a series of studies from the older CVM literature that attempted to test for hypothetical bias in the more complicated setting of public goods valuation in the field.

[2] We generally calculate the hypothetical bias as the excess of the hypothetical to real WTP as a percentage of the real. The exception is when the real WTP is higher than the hypothetical, in which case we report the bias as the excess of the real over the hypothetical as a percentage of the hypothetical.

Bohm (1972) is a landmark study that has had a great impact on many researchers, primarily with respect to the design of field experiments of public goods provision and tests of the extent of free riding and strategic bias. The commodity used in Bohm's experiments was a closed-circuit broadcast of a new Swedish TV program. Six elicitation procedures were used. In each case except one, the TV program was made available and subjects in each group allowed to see it, if aggregate WTP equaled or exceeded a known total cost. Every subject received 50 Swedish Kroner (SEK) when arriving at the experiment.

The question of free riding behavior in public goods provision is relevant to hypothetical bias because part of the latter can be claimed to derive from strategic behavior. In particular, as we discussed in our introductory remarks, hypothetical bias arises from a strategic over-bidding when no actual payment is expected.

Bohm employed five basic procedures for valuing his commodity to capture different degrees and directions of strategic bias. In Procedure I, the subject paid according to his stated WTP. In Procedure II, the subject paid some fraction ($\leqslant 1$) of stated WTP, with the fraction determined equally for all in the group such that total costs are just covered. In Procedure III, subjects did not know the specific payment scheme at the time of their bid, but did know that it was a lottery with equal probability attached to the payment schemes of Procedures I, II, IV, and V. In Procedure IV, each subject paid a fixed amount (SEK 5). In Procedure V, the subject paid nothing. No formal theory was provided to generate free-riding hypotheses for these procedures, and all predictions should be interpreted as making weak inequality predictions. Procedure I was deemed the most likely to generate strategic *under*-bidding, and procedure V the most likely to generate strategic *over*-bidding. The other procedures, with the exception of VI, lay somewhere between these two extremes. Procedure VI was introduced in two stages. The first stage, denoted VI:1, approximates a CVM since nothing was said to the subject about actually being offered the opportunity to watch the program (i.e., it was purely hypothetical). The second stage, VI:2, involved subjects bidding for the right to see the program against what they thought was a group of 100. This was a discriminative auction, with the 10 highest bidders actually paying their bid and being able to see the program.

The major result cited from Bohm's study was that bids were virtually identical for all institutions, averaging between SEK 7.29 and SEK 10.33. Unfortunately, these conclusions are based on parametric test procedures, which are unreliable given the non-Gaussian nature of the samples. The mean contributions for Procedures I–V, VI:1 and VI:2, respectively, were 7.6, 8.8, 7.3, 7.7, 8.4, 10.2, and 10.3 (all in SEK), implying a hypothetical bias that ranges from 16% to 40% (not including VI:2). The respective medians, 5, 8, 5, 5, 7, 10, and 10, in these cases suggest an even larger disparity between the hypothetical institution and the first five procedures: the hypothetical bias ranges from 25% to 100%. Using a non-parametric Kolmogorov–Smirnov test procedure, Cummings and Harrison (1994) derived critical probabilities that Procedure VI:1 elicited the same values as Procedures I–V and VI:2. With the exception of institution VI:2 they conclude that there is a clear presence of hypothetical bias for all institutions.

This means that not all of the hypothetical bias can be explained as strategic bias. These values are included in Table 1.

Kealy, Montgomery, and Dovidio (1990) examine the predictive validity of CVM values for actual cash payment for both a private good (a chocolate bar) and a public good (a de-acidification program for lakes in the Adirondack region). Each subject was first asked for their WTP, and two weeks later the same subjects were asked for an actual payment. 52 out of the 72 respondents to the private good valuation question answered yes, but then given a chance to make an actual purchase only 40 did so. This corresponds to a hypothetical bias of 30%.[3] For the public good, 56 of the 107 respondents said yes hypothetically, but only 28 said yes when given an actual opportunity to contribute. The hypothetical bias is calculated as 100%. Both of these biases are significant.

Another experiment with a public-like good is reported by Seip and Strand (1992). A sample of 101 Norwegians were asked in personal interviews whether they would pay 200 Norwegian Kroner for membership in the Norwegian Association for the Protection of Nature (Norges Naturvernforbund, NNV), which is the largest and best established private environmental organization in Norway. Sixty-four subjects responded "yes." A short time later, the 64 subjects that answered yes in the CVM study were sent letters encouraging them to join the NNV at a membership cost of 200 Kroner. There was no reference in these letters to the earlier CVM study. One month later a second mailing was sent to subjects that had not joined the NNV as a result of the first letter. Again, reference was not made to the initial CVM study. At the end of the second mailing only *six* of the original 64 "yes" respondents in the CVM had actually paid the 200 Kroner to join the NNV. Moreover, all of the hypothetical "no" respondents in the CVM could be reasonably expected to say "no" again if asked for real. Since the expected hypothetical WTP from this sample is 127 Kroner (= 200 Kroner $\times$ 64 yes responses $\div$ 101 possible responses) and the expected real WTP is only 6 Kroner (= 200 Kroner $\times$ 6 yes responses $\div$ 101 possible responses), we conclude that there was a hypothetical bias of 2017%.

Duffield and Patterson (1992) used mail surveys to obtain three sets of values for a fund to be established for the purpose of leasing water rights to be used for the preservation of in-stream flows in a set of Montana rivers. They asked one set of subjects (Cash-TNC) to *actually* make a tax deductible contribution to an *actual* fund, the "Montana Water Leasing Trust Fund," that had been established by the Montana Nature Conservancy. They asked a second group (Hypo-TNC) a hypothetical question: if contacted in the next month with a request to make a tax deductible contribution to the Montana Water Leasing Trust Fund, how much *would* they be willing to contribute?

Apart from finding that the average WTP among respondents was significantly different across the two groups, they also found a very different response rate. The average WTP among respondents to the hypothetical questionnaire was \$12.70 and among

[3] Responses are not reported for each price response, so we have calculated this bias based on the average offer price of \$0.90.

respondents to the actual cash request was $9.40, amounting to a bias of 35%. If non-respondents are included as expressing a zero WTP, the numbers change to $2.97 for the hypothetical and $0.98 for the actual cash and a bias of 203%. In a field study of the introduction of new water systems in India, Griffin et al. (1995) also found some evidence of hypothetical bias. Based on an initial survey in 1988 that asked respondents for their willingness to connect to a potential new system at different rates, they found when re-visiting the respondents in 1991 that 91% of the respondents had, indeed, acted as they claimed they would. Nevertheless, the responses from 1988 over-predicted actual behavior. Among those who claimed in 1988 that they would connect at the relevant rate, 29% never did.

This brief review of results from experiments concerned with the validity of the CVM show us that there is some evidence of hypothetical bias in valuation tasks. Nevertheless, results are sometimes difficult to interpret due to statistical complexities necessitated by the experimental design and the field context in which many of the experiments were conducted. For example, Dickie, Fisher, and Gerking (1987) found that the responses to one of their survey teams was significantly different from the others and that deleting these responses changes the conclusions about hypothetical bias in important ways. Likewise, deleting one respondent who expressed an extremely high demand affects the bias estimate. The interpretation of the results in Bohm (1972) depends on which institution that is selected to represent the true WTP. If institution VI:2 gives the true value, rather than any one of I–V, there is no hypothetical bias. In fact, it could be argued that values in an institution such as VI:2 are underestimates of true WTP, in which case the CVM response is also below the true WTP. The study by Seip and Strand (1992), finally, did a follow-up interview which indicated that responses in the original CVM might not have reflected the question the investigators had posed. Many subjects indicated that the hypothetical value expressed was their WTP for environmental goods in general, and not just for membership in NNV.

Two themes emerge from this brief review:

- There does appear to be some evidence for hypothetical bias in valuation tasks;
- The results can be difficult to interpret and sensitive to variations in experimental design and field conditions.

In partial reaction to these complications and difficulties of interpretation, a series of laboratory experiments were undertaken to identify more directly the extent of hypothetical bias in valuation tasks.

3. Open-ended Elicitation in the Lab

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 an unknown Navajo artist, or a reprint of a medieval map.

The most interesting design feature of these experiments is the attempt to differentiate a generic CVM elicitation from a hypothetical Vickrey auction (HVA). 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 (RVA) treatment except that the instructions were minimally changed to reflect the hypothetical nature of the transaction. The goal of this intermediate design was to see how much of the hypothetical bias might be due to the hypothetical nature of the economic commitment, as revealed by the difference between HVA and RVA, and how much might be due to the absence of a structured institution in a CVM, as revealed by the difference between HVA and the generic CVM.

Their results were clear: the culprit was the lack of a real economic commitment in either of the two hypothetical institutions. Average hypothetical bids for the map were $301 and average real bids were $12, implying a bias of over 2400% between the HVA and the RVA. Between the CVM and the RVA, for the same good, the bias found was 2600%, but for the painting it was less: 290%. This is shown in Table 1. Omitting some outliers, which is common in much of the CVM literature, reduces the bias but does not eliminate it.

McClelland, Schulze, and Coursey (1993) created an interesting hybrid of "induced" and "homegrown" values by generating a market for insurance. Subjects could bid for insurance policies to avoid some low-probability bad outcome that was generated by the experimenters according to a specified probability distribution. To the extent that the risk attitudes of subjects were homegrown, the fact that the underlying risk was induced does not change the fact that the experimenters were eliciting homegrown values. They employed uniform-price auctions in which the top 4 bidders out of 8 would receive the insurance policy at the bid-price of the 5th highest bidder. The uniform-price auction is the multiple-unit analogue of the single-unit Vickrey auction, and shares the same properties in terms of incentives for truth-telling by bidders.

Subjects bid for 10 rounds on each of either 4 or 2 different underlying probability functions, for a total of 40 or 20 bidding rounds. In each case the reigning bids were displayed after each round. In addition, experimenters elicited open-ended hypothetical responses to valuation questions posed both before the session started (referred to as inexperienced responses) and between initializations of new probability functions (referred to as experienced responses). They find that the hypothetical bias changes with the probability function such that at low probabilities hypothetical WTP is about twice that of actual bids, but this bias is reduced, then eliminated, and in some cases reversed, as the probability of a loss increases. The highest bias (120%) was found for inexperienced responses to the lowest risk event. For the two highest risk events the inexperienced responses show that real bids exceed hypothetical by about 25%.

4. Dichotomous Choice Elicitation in the Lab

In response to Neill et al. (1994), many CVM proponents 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. Incentive-compatibility is apparent in DC, at least in the usual partial-equilibrium settings in which such things are discussed.

However, the fact that an institution is incentive compatible when the consequences are real says nothing about the incentive compatibility of the institution when the consequences are not real. Cummings, Harrison, and Rutström (1995) designed some simple experiments, to expose the boldness of the claims that hypothetical DC institutions would be incentive compatible.

Subjects were randomly assigned to one of two treatments, the only difference being the use of hypothetical or real language in the instructions. Both student and non-student adults (drawn from church groups) were employed in this study and both within and between subject treatments were explored. 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." In other sessions subjects valued a box of gourmet chocolate truffles or a small solar-powered calculator. Subjects were asked simply to say whether or not they would be willing to pay some stated amount for the good.

Hypothetical subjects responded much more positively than the real subjects, allowing Cummings, Harrison, and Rutström to reject incentive compatibility. The same qualitative results were obtained with both the student and non-student subjects. Table 1 displays the hypothetical bias found as 163%, 873%, and 163%, for the juicers, chocolates, and calculators, respectively.

Johannesson, Liljas, and Johansson (1998) provide an attempt at understanding the degree to which interpretation lies at the root of the hypothetical bias problem, in replications of the Cummings, Harrison, and Rutström (1995) experimental design. Interpretation was an issue brought up in discussions of the "here and now" wording of the Cummings, Harrison, and Rutström instructions, and also reflected in the follow-up interviews in the Seip and Strand study, discussed above. 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 *only* the *latter* responses as indicating a "yes," they conclude that hypothetical bias disappears. Their results are displayed in Table 1, denoted JLJ.

Smith and Mansfield (1998) employ a DC design in which subjects, who have just participated in a survey, are asked if they would be willing to participate in another, future survey for a given compensation. Five different compensation amounts are used in the design, ranging from $5 up to $50. The primary treatment was whether the compensation was offered in a "real context" or a "hypothetical context." The "hypothetical context" was as follows:

> Researchers at Duke University are considering establishing a sample of households that would be contacted once a year to ask their opinions on programs like the ones I described. This is one of several plans. If this approach were taken and if they could pay you [$ amount] for your continued involvement in two more interviews like the two we have done, would you participate?

The "real context," on the other hand, asked the question as follows:

> Researchers at Duke University are establishing a sample of households that would be contacted again to ask their opinions on programs like the ones I described. They can pay you [$ amount] for your continued involvement in two more interviews like the two we have done. They would send a check for [$ amount] to you when the next two interviews are finished. Will you participate in this new program?

Out of a total sample of 540, roughly split across treatments, 83% of those in the hypothetical context said "yes" to the offer and 82% of those in the real context said "yes" to the offer (their Table III, p. 215). This is not a significant difference.[4] The spread of subjects across DC prices in the two treatments was statistically random (fn. 4, p. 210), so this appears to be strong evidence of the absence of hypothetical bias. Simple statistical models of the DC response provide evidence in support of the hypothesis that the subjects in each treatment used the same choice function when responding.

Two aspects of the experimental design appear as candidates for explaining the lack of a hypothetical bias: the fact that subjects were familiar with the good being valued, having just completed one survey, and the possibility that subjects did not perceive a difference between the hypothetical and the real treatment, given the instructions quoted above.

Frykblom (1997) provides a test of hypothetical bias that involves both open- and closed-ended valuations. The good is a private good, selected because its possible preference relationship to environmental values: an environmental atlas that retails for SEK 200. The real institution is a Vickrey auction. This study found no significant difference between the two hypothetical survey formats. The hypothetical bias based on the open-ended survey is 60%, and it is 56% based on the dichotomous choice survey. Both are significant.

5. Social Elicitation in the Lab

In response to the experimental results of Cummings, Harrison, and Rutström (1995), some CVM proponents argued that their claims for the incentive-compatibility of the DC approach actually pertained to simple majority rule settings in which there was

[4] Lacking information on responses by DC price, we assume the same response rate across all to calculate the hypothetical bias of 3.2% that is reported in Table 1.

some referendum over just two social choices. Somehow that setting would provide the context that subjects need to spot the incentive compatibility, or so it was argued.

In response to these arguments, Cummings et al. (1997) undertook 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.

The key treatment in their simple experimental design was again the use of hypothetical or real payments, and again there was significant evidence of hypothetical bias. The hypothetical bias found is 67%, and it is significant. Table 1 shows this.

6. Constructive Solutions

There have been two broad, constructive responses to the evidence of the existence of hypothetical bias. Each entails an effort to use experimental methods to calibrate the extent and the correlates of hypothetical bias. We refer to these as "instrument calibration" and "statistical calibration."

6.1. 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. It can provide 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. The problem of deciding which set of words is "best" might, in some instances, be easily and directly tested using controlled laboratory experiments. Unfortunately, to our knowledge no published experimental results are available that test the notion of instrument calibration. Several unpublished studies have been circulated, however, such as Cummings, Harrison, and Osborne (1995) and Cummings and Taylor (1999).

6.2. Statistical Calibration

The word "calibrate" is defined in *The Oxford Dictionary* as to "determine or correct the calibre or scale of a thermometer/gauge or other graduated instrument." 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 has been offered by Blackburn, Harrison, and Rutström (1994), Fox et al. (1998) and Harrison et al. (1996). 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. Blackburn, Harrison, and Rutström (1994) offer 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 is *transferable* to other instances (i.e., the watch does not get further behind the times over time). The trick is how to estimate and apply such bias functions. This is done with the *complementary* use of 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.

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 be predicted to 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."[5]

The statistical calibration approach can do more than simply point 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 well 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.

[5] Mitchell and Carson (1989) provide a popular and detailed review of many of the traits of "bad hypothetical surveys."

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.

This simple test was encouraging, but was limited by design to deliverable private goods such as juicers and chocolates. Harrison et al. (1998) attempted to generalize their procedures to the case of greater interest in field applications of hypothetical surveys to assess environmental damages. They undertook five surveys, each one designed to provide information that would allow for the calibration of a field public good. One survey was for a deliverable private good, and used a Vickrey auction with real payment required. The next survey varied this by allowing for public provision of the private good, albeit with real payment required. Comparison of the results of the two surveys, after correcting for socioeconomic differences in the sample, was intended to provide a statistical measure of the propensity for free-riding bias. Similarly paired surveys provided a measure of the propensity for hypothetical bias in the provision of a deliverable public good. Arguing that the propensity to engage in hypothetical bias would be independent of the propensity to engage in free-riding bias, the results from the initial two surveys were then used to adjust the results in the latter two surveys used to elicit hypothetical bias. The end-result was a statistical measure of the propensity of subjects to engage in both types of bias. This measure was used to adjust hypothetical responses to a survey asking for valuation of a non-deliverable public good.

Although the statistical procedures employed by Harrison et al. (1998) were illustrative, their design demonstrates the potential complementarity between lab and field experiments. Their results were surprisingly encouraging. The most important result is that they cannot reject the hypothesis that true WTP is equal to the raw WTP elicited by a hypothetical survey in the case of national wetlands preservation (their non-deliverable field public good). However, they can reject this hypothesis in the case of local wetlands preservation (their deliverable field public good). These results illustrate well the view that one can validate *or* calibrate the results from CVM surveys in a constructive manner. If the CVM result is a valid reflection of true WTP, calibration methods can show that it is (as in the case of national wetlands preservation). However, if the CVM result is not a valid reflection of true WTP, calibration methods not only show that, but are able to provide some guide as to how serious the bias is.

Fox et al. (1998) discuss and apply a calibration approach to hypothetical survey values that uses experiments to ascertain the possible hypothetical bias. Their approach differs from BHR and Harrison et al. (1998) by using experiments to calibrate hypothetical to real values for the same good, rather than using experiments to determine some calibration function that can be used to calibrate across different goods. One setting in

which it could be practically useful is in which one conducts a large-scale hypothetical survey and seeks to calibrate it with smaller-scale real experiments.

The commodity used was health risk reduction in a food product. Telephone interviews with 174 pork-eating respondents selected at random in Story County, Iowa, resulted in a series of hypothetical valuations. Of these, 75% indicated a preference for an irradiated food product over the "raw" food product. Each subject was also asked to state an open-ended hypothetical valuation for their preferred product.

From these subjects, 78 agreed to participate further in some lab experiments in return for a payment of $30. Those subjects were endowed with their less-preferred food product from the hypothetical telephone survey, and then a series of auctions were undertaken to see what they would be willing to pay to "upgrade" to their preferred food product. Each subject was reminded, before bidding, of their previous hypothetical valuation and that this bid would be used as their initial bid in round 1. After 5 rounds of bidding, subjects were given information on the difference between the two products, and allowed to bid for 5 more rounds. One of the rounds was chosen at random, and the transactions effected. In one treatment group there was only one bidding round, and in some treatments the standard single-unit Vickrey auction rules were modified to allow a random nth price auction.[6] The potentially winning bids and the ID numbers of the winning bidders was announced after each round.

The results suggest calibration factors of roughly two-thirds if one compares hypothetical survey values and the round 2 auction values. The auction values in the final round are generally higher than those in round 2, so the calibration factors are higher as well (between 60% and 83%).

7. Conclusions

There seems to be little doubt that the presumptive absence of hypothetical bias in CVM surveys is invalid. It is invalid as a general proposition, since the experiments surveyed here provide special cases when it is demonstrably false. The importance of these experimental results is that they change the tenor of the debate on the validity of the CVM. The variety of elicitation formats, the variety of subject pools, and the variety of private and public goods all serve to show that one cannot ignore the problem. Of course, it is possible that hypothetical bias may be absent or swamped by other biases in *particular* settings. But one cannot make sweeping and unqualified claims as to its absence, as has been common in the CVM literature.

There is some experimental evidence that comparing simple "yes" responses in real settings to "definitely yes" responses in hypothetical settings indicates less hypothetical bias than when simple "yes" responses in hypothetical settings are used. Although these

[6] If N is the number of bidders, then a number n between 1 and $N-1$ is chosen by the auctioneer. Everyone who bids greater than the bid of the nth bidder wins and pays the bid of the nth highest bidder.

results provide some promise for generating hypothetical responses that better match real responses, no theory underlies these results and no guarantees can be made that hypothetical bias is removed in all settings with this change in interpretation.

There is some evidence that the extent of hypothetical bias varies from setting to setting, but the sample sizes and design employed thus far are too slight for one to draw any broad conclusions from that. Hence it is particularly inappropriate to try to claim that hypothetical bias is any more of a problem in open ended formats as compared to closed ended formats, or that referendum formats are "less inaccurate" than DC formats.

Attempts at calibrating responses of hypothetical surveys have been reasonably successful. This implies that experimental methods may well have an important role to play in improving the reliability of hypothetical survey methods.

References

Bishop, Richard, Heberlein, Thomas (1979). "Measuring values of extra market goods: Are indirect measures biased?" American Journal of Agricultural Economics 61, 926–930. [BH].

Bishop, Richard, Heberlein, Thomas (1986). "Does contingent valuation work?" In: Cummings, R., Brookshire, D., Schulze, W. (Eds.), Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Valuation Method. Rowman & Allenheld, Totowa, NJ.

Bishop, Richard C., Heberlein, Thomas A., Kealy, Mary Jo (1983). "Contingent valuation of environmental assets: Comparisons with a simulated market". Natural Resources Journal 23, 619–633. [BHK].

Blackburn, McKinley, Harrison, Glenn W., Rutström, E. Elisabet (1994). "Statistical bias functions and informative hypothetical surveys". American Journal of Agricultural Economics 76 (5), 1084–1088.

Bohm, Peter (1972). "Estimating the demand for public goods: An experiment". European Economic Review 3, 111–130.

Cummings, Ronald G., Harrison, Glenn W. (1994). "Was the *Ohio* Court well informed in their assessment of the accuracy of the contingent valuation method?" Natural Resources Journal 34 (1), 1–36.

Cummings, Ronald G., Taylor, Laura O. (1999). "Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method". American Economic Review 89, 649–665.

Cummings, Ronald G., Harrison, Glenn W., Osborne, Laura L. (1995). "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.

Cummings, Ronald G., Harrison, Glenn W., Rutström, E. Elisabet (1995). "Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive compatible?" American Economic Review 85 (1), 260–266. [CHR].

Cummings, Ronald G., Elliott, Steven, Harrison, Glenn W., Murphy, James (1997). "Are hypothetical referenda incentive compatible?" Journal of Political Economy 105 (3), 609–621. [CEHM].

Dickie, M., Fisher, A., Gerking, S. (1987). "Market transactions and hypothetical demand data: A comparative study". Journal of the American Statistical Association 82, 69–75. [DFG].

Duffield, John, Patterson, David A. (1992). "Field testing existence values: An instream flow trust fund for Montana rivers". National Center for Environmental Economics Report EE-0282.

Fox, John A., Shogren, Jason F., Hayes, Dermot J., Kliebenstein, James B. (1998). "CVM-X: Calibrating contingent values with experimental auction markets". American Journal of Agricultural Economics 80, 455–465.

Frykblom, Peter (1997). "Hypothetical question modes and real willingness to pay". Journal of Environmental Economics and Management 34 (3), 275–287.

Griffin, Charles C., Briscoe, John, Singh, Bhanwar, Ramasubban, Radhika, Bhatia, Ramesh (1995). "Contingent valuation and actual behavior: Predicting connections to new water systems in the state of Kerala, India". The World Bank Economic Review 9 (3), 373–395.

Hanemann, W. Michael (1984). "Welfare evaluations in contingent valuation experiments with discrete responses". American Journal of Agricultural Economics 66, 332–341.

Harrison, Glenn W., Beekman, Robert L., Brown, Lloyd B., Clements, Leianne A., McDaniel, Tanga M., Odom, Sherry L., Williams, Melonie (1998). "Environmental damage assessment with hypothetical surveys: The calibration approach". In: Boman, M., Brännlund, R., Kriström, B. (Eds.), Topics in Environmental Economics. Kluwer Academic Press, Amsterdam.

Johannesson, Magnus, Liljas, Bengt, Johansson, Per-Olov (1998). "An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions". Applied Economics 30, 643–647. [JLJ].

Kealy, M.J., Montgomery, M., Dovidio, J.F. (1990). "Reliability and predictive validity of contingent values: Does the nature of the good matter?" Journal of Environmental Economics and Management 19, 244–263. [KMD].

McClelland, Gary H., Schulze, William D., Coursey, Don L. (1993). "Insurance for low-probability hazards: A bimodal response to unlikely events". Journal of Risk and Uncertainty 7, 95–116. [MSC].

Mitchell, Robert C., Carson, Richard T. (1989). "Using Surveys to Value Public Goods: The Contingent Valuation Method". Johns Hopkins Press, Baltimore.

Neill, Helen R., Cummings, Ronald G., Ganderton, Philip T., Harrison, Glenn W., McGuckin, Thomas (1994). "Hypothetical surveys and real economic commitments". Land Economics 70 (2), 145–154.

Seip, K., Strand, J. (1992). "Willingness to pay for environmental goods in Norway: A contingent valuation study with real payment". Environmental and Resource Economics 2, 91–106. [SS].

Smith, V. Kerry, Mansfield, Carol (1998). "Buying time: Real and hypothetical offers". Journal of Environmental Economics and Management 36, 209–224. [SM].