Values Elicited from Open-Ended Real Experiments

by

J.K. Horowitz and K.E. McConnell

WP 98-10

Department of Agricultural and Resource Economics
The University of Maryland, College Park

VALUES ELICITED FROM OPEN-ENDED REAL EXPERIMENTS

by

John K. Horowitz and K.E. McConnell

Department of Agricultural and Resource Economics
University of Maryland
College Park MD 20742-5535

February 1999

Send correspondence to:
John K. Horowitz
Department of Agricultural and Resource Economics
University of Maryland
College Park MD 20742-5535
(301) 405-1273
(301) 314-9091 (fax)
horowitz@arec.umd.edu

VALUES ELICITED FROM OPEN-ENDED REAL EXPERIMENTS

Economists frequently conduct experiments in which public or private goods are hypothetically traded for money. The responses to these experiments are taken to be motivated by preferences, and so the experiments are useful in learning about preferences and valuation. For pure public goods, hypothetical experiments such as contingent valuation are sometimes the only way to learn about preferences. The results of hypothetical experiments help determine the policies economists will recommend for environmental amenities and other public goods.

The idea that preferences are only revealed by real incentives is deeply embedded in economists' worldview. Consequently, evidence from hypothetical experiments has not readily permeated economic thinking. Students of hypothetical valuation have argued that one method for determining whether these hypothetical experiments provide useful information about preferences is to compare them to similar real experiments in which goods and money are traded in a controlled environment. Because real experiments involve real incentives, the argument goes, they provide unimpeachable evidence about preferences. If otherwise similar real and hypothetical experiments yield similar results about preferences, one can conclude that hypothetical experiments provide useful information for policy. The reliability of evidence from real experiments is a critical link in this conclusion.

The role played by real experiments in this link presumes that the 'realness' of the experiments, induced by exchanges involving money, elicits preferences. In this paper, we argue that results from a set of real experiments do not conform to what we would expect for real preferences. This raises the question of whether experimental

results, real or hypothetical, constitute a useful guide for policy.

Another interpretation of the results relates to the criticisms often applied to contingent valuation. Our real experiments show that several kinds of non-economic behavior are exhibited. One can turn this evidence around to argue that when contingent valuation experiments produce odd or non-economic results, the results are not different from what one finds in real experiments, and probably in real social-choice behavior.

The paper approaches the question of the validity of the experimental responses by applying a set of criteria that range from a broad appeal of implausibility to a narrow restriction based on quasi-concavity of preferences. The analysis examines the proportion of respondents from a real experiment that meet each of the criteria. We will argue that these proportions are unreasonably low. We conclude that the link from hypothetical experiments to policy conclusions presumed to be provided by real experiments is weak.

1. The Experiments

Three experiments were conducted in-person with small groups. Subjects were endowed either with 3 pairs of binoculars, 3 flashlights, or 3 mugs. The experiments differ from others available in the literature by one or more of the following features: (i) Multiple items of the same good were valued. (ii) The market value of the items was relatively high, especially in the binocular experiment, where the items being valued cost us $75. (iii) The questions were posed as willingness-to-accept (also referred to as compensation demanded, CD) rather than willingness-to-pay. (iv) The

subjects were members of the public. (v) A Becker-DeGroot-Marschak (BDM) mechanism with an unspecified range of offer prices was used to elicit compensations demanded.

The binocular experiment was performed as follows. Flashlight and mug experiments were similar. Each subject was given one mug and three pairs of binoculars. We first asked subjects to consider selling back the mug. Each participant was asked to write down the minimum payment he or she would require to be willing to sell the mug back to us; this is his compensation demanded. We then repeated the following procedure four times, three times for practice and then for real money and a real transaction. The administrator drew an *offer price* randomly out of an envelope. If the subject's compensation demanded was higher than the offer price, the subject kept his mug. If his compensation demanded was less than or equal to the offer price, he returned his mug to us and received a check for the randomly drawn price. All subjects were offered the same offer price. In the flashlight experiment, the subjects were given 1 mug and 3 flashlights. In the mug experiment, the subjects were given 1 flashlight and 3 mugs. We do not study the values elicited in this first part, which was used to teach the price mechanism to subjects. This mechanism is called a BDM mechanism.

In the second part of the experiment, each subject wrote down two numbers: the minimum payments he or she required to be willing to sell us back two and three binoculars. These are the subject's compensation demanded for two and three binoculars. We then randomly drew a piece of paper that stated the number of binoculars (per person) we would be buying back. Each option had equal probability, and subjects

were told this. For binoculars and mugs, the probability was ½ that we would offer to buy back 2 items and ½ to buy back 3. For flashlights, the probability was $^1/_3$ that we would offer to buy back 1, 2, and 3 items.

We then randomly drew the offer price. For example, we might randomly draw the instruction to buy back two of each subject's three binoculars, and then draw an offer price of $19.00. This is a price for the two binoculars, not a per-binocular price. All subjects who had offered to sell two binoculars for $19.00 or less then turned in two of their binoculars and received a check for $19.00. They kept their remaining pair of binoculars. All subjects who had offered to sell two binoculars for more than $19.00 kept all three binoculars and received no money.

Subjects were not told the distribution of offer prices, but they knew that separate distributions were used for the two and three binocular cases. Once the number of binoculars at stake was chosen, the offer-price mechanism was the same as for the single mug, except that this part of the experiment was conducted only once, for real money and a real transaction.

Our decision not to reveal the distribution of offer prices makes our mechanism different from the BDM mechanism as it is most frequently administered. In particular, we did not tell subjects the upper limit of the distribution, *i.e.*, the highest potential offer price. Concealing this information was useful to us for two reasons. First, it further emphasizes to subjects that distribution information should be irrelevant in their responses; that is, it helps remind them that their best strategy is to determine and report their compensation demanded regardless of what they believe about the possible offer prices. Second, it gives the subjects no information about what anyone else

(namely, the experimenters) believes might be likely values for CD. The latter feature is obviously important in studying "real values."

Subjects were members of local civic groups. The sponsoring group was paid $20 per subject. (Payments for returning the items went directly to the subjects.) Local civic groups are more representative of the U.S. population than other commonly surveyed groups such as students. The cost per response is considerably less than an in-person survey of a more rigorously selected sample. The groups are not representative samples of the U.S. population but they are probably not systematically different from the respondents in most other non-student valuation studies.

2. Criteria

We are interested in the values elicited by these experiments. Before one can examine those values, some responses may be deleted for being faulty in one way or another. We set out four criteria to show how the experiments perform under different samples. Some responses are clearly wrong in the sense that one can be reasonably assured that a respondent confronted with his responses would admit they were not his true preferences. Others are wrong for more subtle reasons. The criteria we propose cover a variety of reasons for considering an observation suitable. For each of the experiments, we ask whether the responses meet the following four criteria.

a) Intuitive plausibility. In the experiments reported, some of the bids are implausible, in the sense that almost all observers would agree that they do not represent preferences that any subject would profess to have. In other words, it would seem to be relatively simple to design a comparable experiment in which the subject

expressed the opposite preference. These responses have typically been called outliers. For example, one subject reported his CD for 2 binoculars to be $10 million.

b) Economic plausibility. When respondents assess their compensation demanded for market goods, they ought to bear in mind the resale value and replacement cost of the goods; in other words, the opportunity costs of otherwise buying or selling the items. Let CD^i represent the subject's compensation demanded to return i items. CD^i ought not to exceed the purchase price of those i items, plus transaction cost, because the respondent can replace any or all of his items if offered enough money. CD^i ought to exceed resale value because the subject can resell the goods if he gets little utility from having them. Hence, CD^i ought to be bounded as follows:

(1) Resale value of i items $\leq$ CD for i items $\leq$ Replacement cost of i items

Economic plausibility requires these bound be met. We put the resale value at zero.

c) The third criterion is based on the marginal compensation demanded rather than total CD. Marginal CD is the extra compensation demanded to give up the ith item, $MCD = CD^i - CD^{i-1}$, with $CD^0 = 0$.

Marginal CD ought to be higher than the resale value of one item and lower than the replacement cost. Suppose for some non-marketed good, a subject's CD for 2 items (when he started with 3) was $6 and for 3 items was $14. Consider what happens if the item could be replaced for $5. The subject's compensation demanded for 3 items should now be no more than $11, because for $11 he could buy one of the items and have $6 left over, which gives him the same commodity-money outcome as in the two-item case. Hence, MCD ought to be bounded as follows:

(2) Resale value per item ≤ Marginal CD ≤ Replacement cost per item

If the subject violates the economic plausibility criterion, which is based on total value, she must violate this marginal criterion. But she can violate the marginal criterion without violating the total criterion. We break the marginal criterion into two criteria based on the first and second inequalities:

c′) Positive marginal value. When the resale value per item is zero, this criterion is identical to requiring that subjects have a positive marginal value for the item: MCD > 0. In other words, respondents ought to require more compensation for more goods. When a respondent asks for the same or less compensation for more goods, the criterion fails. This criterion is equivalent to a within-subject test for scale sensitivity in a contingent valuation context.

c″) Bounded marginal value. This criterion is the second inequality in expression (2).

d) Quasi-concavity of preferences. When multiple items are valued, it is possible to trace out the curvature of the indifference curve between money and the good. If the marginal compensation demanded is increasing in the number of items given up, then the indifference curve would have the usual shape in money and goods. Strictly increasing MCD means that CD^i responses are strictly convex in *i*. Strict convexity is the standard assumption of neoclassical economics. It predicts the pattern {CD^1 = \$2, CD^2 = \$5, CD^3 = \$9} but not {CD^1 = \$4, CD^2 = \$7, CD^3 = \$9}. Both patterns meet all of the preceding criteria for a good with replacement cost of \$5.

When a subject is asked to report CD for giving up one, two, and three items, as in the flashlight experiment, there are four ways to determine convexity: $2CD^3 > 3CD^2$,

$CD^3 > 3CD^1$, $CD^2 > 2CD^1$, and $CD^3+CD^1 > 2CD^2$. For experiments when subjects report only CD for giving up two and three items, only the first expression can be evaluated. If CD^i is strictly convex in i (the standard assumption), each of these expressions will be positive. These tests are not always independent. But if any fails, then the subject has preferences that are not strictly quasi-concave. (See Horowitz, McConnell, and Quiggin for further analysis of convexity.)

Failure of quasi-concavity is different from failure of the other criteria. Respondents can have transitive and believable preferences and be fully exploiting any available opportunity costs, yet still fail quasi-concavity.

For market goods, the bounded marginal value criterion confounds observation of violations of strict quasi-concavity. When replacement cost is \$5, a subject might report $\{CD^1 = \$4, CD^2 = \$9, CD^3 = \$14\}$ even if he had strictly quasi-concave preferences; he might have had higher compensation demanded for his third item if a market replacement were not available, but because he can replace his third item on the market, he reports $CD^3 = CD^2$ plus replacement cost.

The combined effect of c″ and d is to allow strictly increasing followed by constant MCD; everywhere constant MCD; or strictly increasing MCD, when c″ is not binding, but not any other patterns. The combined effect of c′ and d is to reverse these patterns; in other words, MCD might be constant for a while, then increasing. For example, a subject might report $\{CD^1 = \$2, CD^2 = \$4, CD^3 = \$8\}$ for a market good whose resale value is \$2. In either of these cases, however, MCD should be constant only when it is equal to the resale value or the replacement cost, and in no case will MCD decrease. All of these patterns are accommodated by preferences that are weakly, rather than strictly, quasi-concave.

Criteria c and d require at least two valuation questions. They would not be usable if only the value for one set of items were elicited. Criteria a and b can be applied to all valuation experiments.

3. Results

A. Binoculars

The 48 responses for the binocular experiment are shown in Table 1. The wholesale price for the binoculars was $25 per pair.

a) Intuitive plausibility. The five highest observations for 2 binoculars were $250, $300, $1000, $10 million, and $20 million. Designating $10 million and $20 million as implausible is uncontroversial. Decisions about $250, $300, and $1000 are not so easily made. Subject 46, whose CD was $1000, would almost surely have said, in some other context, that he preferred $975 rather than 2 pair of binoculars, and so we treat his responses as intuitively implausible. We therefore take the three highest responses to be intuitively implausible, leaving 45 observations.

b) Economic plausibility. Binoculars like the ones used in the experiment cost us $25. We suggest that replacement cost is less than double our cost of $50 for two pair of binoculars and $75 for three pair.[1] At these payoffs, respondents could easily purchase equivalent binoculars, cover any transactions costs, and pocket some extra money. We designate observations 39 through 45 as economically implausible. This leaves 38 of the original 48 observations.

[1]Although wholesale prices are typically substantially below retail prices, which is what most subjects would have to pay for these items, we did not "shop around" to get the lowest wholesale price. With relatively little effort, subjects would likely be able to buy similar items for about what we paid for them.

c') Positive marginal values. A positive marginal value fails if $CD^2 \geq CD^3$, since these subjects want less than or the same compensation for three binoculars as for two. This rule further eliminates observations 4, 8, and 22. There are no remaining observations that fail the bounded marginal value criterion. (Observations 39, 40, and 43, which fail the economic plausibility test, also fail to have a positive marginal value.) This leaves 35 of the original 48 observations.

d) Quasi-concavity of preferences. Of the 35 observations passing the first three tests, 17 have values that are linear in the number of binoculars relinquished, 14 are strictly convex (the typical utility-theoretic prediction), and 4 are strictly concave (nos. 2, 27, 33 and 38). The strict convexity criterion leaves 14 of 48 original observations. The weak convexity criterion leaves 31 of 48 observations.

The two highest CD's conform to standard utility-theoretic predictions because $CD^3 - CD^2 > \frac{1}{2}CD^2$. The \$1000 response satisfies this inequality weakly. In other words, the intuitively implausible responses are well-behaved under this criterion.

B. Flashlights

The flashlight experiment differed from the binocular experiment by (i) eliciting subjects' compensation demanded to give up one item of their endowment, as well as two and three items; and (ii) giving subjects an explicit option for declining to give back any flashlights. The 42 responses for flashlights are shown in Table 2.

a) Intuitive plausibility. Two respondents declined to sell any of their flashlights at any price, a decision that is intuitively implausible. Three additional respondents declined to sell their third flashlight at any price. This criterion leaves 37 of the 42 original flashlight observations.

b) Economic plausibility. We take replacement cost to be less than double our cost of $6.25 per flashlight. Observations 36, 38, 39, and 40 are eliminated this way. This criterion leaves 33 of the 42 original observations.

c′) Positive marginal values. Subjects can reveal non-positive marginal values if $CD^1 \geq CD^2$ or if $CD^2 \geq CD^3$. Four respondents reveal non-positive marginal values (nos. 1, 12, 22, 37). This criterion leaves 29 observations.

c″) Bounded marginal value. Two remaining respondents reveal marginal values for the third flashlight that exceed the replacement cost bound of $12.50 (nos. 19 and 32).

d) Quasi-concavity of preferences. Of the 27 remaining observations, the strict quasi-concavity criterion leaves 2 observations (nos. 6 and 7). The weak quasi-concavity criterion leaves 11 observations; the previous two plus 7 subjects whose valuations are linear (nos. 4, 8, 13, 14, 23, 31, 35) and 2 whose valuations are linear then convex (nos. 3, 27).

The flashlight experiment differed in asking for compensation for just one item. This change might be expected to encourage linear responses. However, only seven observations in the 37 observation set are completely linear.

C. Mugs

The 41 responses for the mug experiment are in Table 3. As in the flashlight experiment, subjects were given an explicit option for declining to give back any mugs.

a) Intuitive plausibility. The compensation demanded ranged from $2 to $20 for two mugs and from $2.50 to $28.50 for three mugs. The extremes are not obviously

implausible. Further, no subject stated he or she would refuse to sell a mug at any price. No observations are excluded for this reason.

b) Economic plausibility. Mugs like the ones used for this experiment are widely available for around $5-$7. We put the replacement cost at just below $8 per mug. Therefore, we view the values revealed by observations 37 through 41 as economically implausible. This leaves 36 observations of the original 41 observations.

c′) Positive marginal values. All of the 36 remaining subjects asked more for three mugs than two. No observations are removed for non-positive marginal values.

c″) Bounded marginal value. One subject revealed a marginal value for the third mug that exceeded the replacement cost bound (no. 14).

d) Quasi-concavity of preferences. Of the 35 remaining observations, 6 have values that are strictly convex in the number of mugs relinquished (nos. 4, 5, 15, 16, 17, and 20), and 17 have values that are linear.

4. Analysis

A. Effects of Criteria on the Distribution of Values

The common goal of hypothetical and real valuation experiments is the estimation of a measure of central tendency of willingness to pay or compensation demanded. It is clear that deleting observations because they are too high will reduce the mean and median of the remaining observations. Less clear, however, is the impact on other sample characteristics, such as dispersion and skewness. Further, it is not obvious how restricting the sample to observations with positive marginal values or convex responses will influence the central tendency. In Tables 4, 5, and 6 we explore

the impact of increasingly restrictive samples. In these tables the sample statistics are calculated with different samples, labeled a, b, c′, c″, and d. In sample a, the intuitively implausible observations are deleted; in b, the economically and intuitively implausible are further deleted; in c′ non-positive marginal value observations; in c″ the observations with too-high marginal values; and in d, valuations that are not consistent with weakly quasi-concave preferences.[2] We look at the weakest of possible restrictions for sample d.

Table 4. Summary Statistics for Binoculars CD (N = 48)

	Criterion	Mean	Coeff. of Variation	Median	Skewness	Percent of Original
CD for 2 binoculars	a	$50.32	1.26	$29.00	2.62	94%
	b	$27.36	0.60	$21.38	1.04	79%
	c′	$28.27	0.59	$22.50	0.95	73%
	d	$26.68	0.53	$20.25	0.52	65%
CD for 3 binoculars	a	$72.63	1.27	$45.00	2.88	94%
	b	$41.40	0.58	$37.25	0.60	79%
	c′	$43.66	0.54	$40.00	0.52	73%
	d	$42.11	0.50	$39.50	0.35	65%

[2]For a and b, we remove CD^i responses only if the response is implausible for *i* items. In the previous section, we counted a subject only if his responses met the criterion for all *i*. For c and d in the flashlight case, we remove the entire response if it violates a marginal or quasi-concavity condition anywhere.

Table 5. Summary Statistics for Flashlights CD (N = 42)

	Criterion	Mean	Coeff. of Variation	Median	Skewness	Percent of Original
CD for 1 flashlight	a	$6.19	0.65	$5.00	1.54	95%
	b	$5.38	0.52	$5.00	0.79	88%
	c′	$5.38	0.50	$5.00	0.78	79%
	c″	$5.29	0.49	$5.00	0.82	74%
	d	$4.61	0.55	$4.00	1.20	26%
CD for 2 flashlights	a	$10.82	0.67	$9.25	1.94	95%
	b	$9.24	0.49	$8.50	0.70	88%
	c′	$9.74	0.45	$9.00	0.75	79%
	c″	$9.50	0.45	$8.50	0.85	74%
	d	$9.68	0.51	$8.50	1.05	26%
CD for 3 flashlights	a	$16.79	0.67	$14.00	1.03	88%
	b	$13.89	0.56	$12.00	0.68	79%
	c′	$14.82	0.50	$12.75	0.82	69%
	c″	$13.69	0.46	$12.00	0.86	64%
	d	$15.84	0.45	$15.00	0.73	26%

Table 6. Summary Statistics for Mugs CD (N = 41)

	Criterion	Mean	Coeff. of Variation	Median	Skewness	Percent of Original
CD for 2 mugs	a	$9.18	0.53	$8.50	0.54	100%
	b	$8.17	0.48	$8.00	0.25	90%
	c″	$8.23	0.48	$8.00	0.20	88%
	d	$7.25	0.46	$7.00	0.26	59%
CD for 3 mugs	a	$13.33	0.49	$14.00	0.29	100%
	b	$12.32	0.45	$13.24	-0.02	93%
	c′	$12.11	0.45	$12.97	0.02	90%
	c″	$12.03	0.46	$12.49	0.06	88%
	d	$11.13	0.45	$11.00	0.10	59%

The number of observations that are economically plausible but fail to have positive marginal values and weakly quasi-concave preferences ranges from 14% to 69% of the full sample, but the effect of the last two criteria on the central tendency

and dispersion of the distributions is mixed. The means change little and the measures of dispersion appear to decline slightly. The main effect of exclusions c and d is the loss of precision because of the reduction in observations.

B. Distribution of Values - Only Outliers Removed

Real and hypothetical experimenters commonly analyze responses without appealing to criteria b, c, or d, either because the criteria cannot be applied (*e.g.*, only one set of items is valued, which eliminates the marginal value and convexity criteria; or the item is not a private good, which nearly eliminates the economic plausibility criterion) or because the experimenters accept the reported values without applying any judgment. In both of these cases, however, analysts are likely to remove obvious outliers. Table 7 presents statistics from all three experiments with only the intuitively implausible responses deleted.

Table 7. Summary Statistics for All Observations (excluding outliers)

	Item		
	Binoculars (per pair)	Flashlights (per flashlight)	Mugs (per mug)
Mean	$24.69	$5.78	$4.52
Coefficient of Variation	1.26	0.65	0.51
Median/Mean	0.61	0.87	1.00
Maximum/Mean	6.08	3.45	2.21
Skewness	2.70	1.50	0.44

Lognormal Distributions. The lower the mean CD, the smaller the upper tail of the distribution of responses. We find that a lower mean response is accompanied by: (a) a lower standard deviation, relative to the mean; in other words, a lower coefficient of

variation; (b) a median closer to the mean; (c) a lower maximum value, relative to the mean; and (d) a lower skewness. These findings are consistent with value distributions that are roughly *lognormal*, conforming approximately to the shape of a lognormal although not to its range.

Number of Outliers. The lower the cost of the item, the lower is the number of outliers that are removed.

Market Value and Mean CD. When only outliers are removed, mean CD's are remarkably close to the item cost of $25 (binoculars), $6.25 (flashlight), and $4 (mug). A brief look at willingness-to-pay (WTP) experiments shows an analogous pattern. Empirically, mean WTP is roughly half the market price, say 40 to 60 percent. For example, Loomis *et al.* elicited willingness to pay for an art print using a first-price auction. Their third treatment, which is the closest to our experiments, found a mean WTP of $14.48, which is 41% of the item cost of $35. Johannesson, Liljas, and O'Conor used a second-price WTP auction for chocolates with 10 participants. The cost was 150 Swedish crowns and the mean WTP was 87.40, which is 58% of the price. Neill *et al.* conducted a second-price WTP auction for a map. Mean WTP was either 50% or 60% of the $20 cost, depending on the treatment of outliers. Horowitz and McConnell reviewed studies that collected both CD and WTP and found that for ordinary private goods, CD was roughly 2.3 times as large as WTP. If CD is close to market value, then WTP will be 43% of that value.[3]

These results – that mean CD values are close to market value – are striking, but they further indicate how much of the aggregate reported value, roughly half, exceeds the replacement-cost upper bound on CD imposed by economic theory.

5. Interpreting the Results

Intuitively Implausible Responses. By far the most important decision for analysis is the elimination of implausibly high responses. The effect of outliers is most obvious for the binocular experiment but the effect is also substantial, although the decision less arbitrary, in the flashlight experiments, where five subjects stated that they were unwilling to return some of their flashlights for any offer price. These responses are ones that subjects themselves would likely admit did not truly reflect their preferences. Low outliers, in which subjects report a value that they would prefer not to accept if it were offered in some other context, may also be present, although they are more difficult to detect and therefore have not been analyzed here.[4]

We believe the outliers cannot be explained based on the criticism often made of valuation surveys in general, that the survey places too high cognitive demands on subjects. Subjects might react to our experiment by reporting high values that guaranteed they keep their items, especially in the binoculars experiment. To guarantee this, a subject needs to guess the highest possible offer price and then state a higher value. (It is not clear that a subject's guessing the highest offer price places lower cognitive demands than deciding the lowest value he would accept.) A reasonable supposition is that the highest offer price for 3 binoculars is 1½ times the highest offer for 2 binoculars. Responses would follow this pattern as well. Only one of the outliers conformed to this strategy. An alternative response strategy is to guess the highest possible offer price over both 2 and 3 items and report the same high value for both cases. None of the outliers

[3]Neill *et al.* also conducted a second-price WTP auction for an art print. Mean WTP was $9.49, which is 13% of the $75 cost, which deviates from the general pattern. Art prints may be too unique to count reliably as ordinary private goods.

[4]In other experiments we conducted similar to these, subjects have reported CD's of below $1.00.

conformed to this approach.

It is possible that intuitively implausible responses will be a problem only for CD experiments, in which subjects are endowed with some item and offered money to return it. CD experiments are much less common than WTP experiments, primarily because response amounts and refusal rates are typically unacceptably high.[5] However, if real experiments are therefore unreliable for inferring preferences based on CD, then they cannot be reliable for WTP, since the same economic models are used to motivate and derive lessons from both.

Positive Marginal Values. The failure for responses to reflect positive marginal values is one of the most unexpected aspects of the experiments. One explanation is that subjects misunderstood the instructions and reported per-item values even though verbal instructions, as well as the written survey, repeatedly emphasized that the subjects should give the total value, not per-item value. A second explanation is that subjects wanted to guarantee the amount of their payment, conditional on receiving payment. This is analogous to the outlier explanation (subjects wanted to guarantee they keep their binoculars), except in the money dimension rather than the goods dimension.

Zero marginal value may occur because of satiation, which seems most likely to happen with mugs. All of the plausible mug responses had positive marginal values.

The anomaly of positive marginal values is akin to a problem in contingent valuation, called insensitivity to scale, in which respondents report approximately the same values for packages of different sizes. A second contingent valuation problem, called insensitivity to scope, arises when subjects report approximately the same values for packages with different attributes, for example two pairs of binoculars versus two

[5]See Neill *et al.* for a recent example in which outliers are a potential problem for WTP experiments.

flashlights versus two mugs. In our experiments, responses are clearly sensitive to scope.

Reasonable Marginal Values. The possibility that marginal values might instead be too *high* even when the total value is reasonable has not be recognized in the valuation literature. We found one or two subjects who violated this criterion for the mugs and flashlights experiments.

Quasi-Concavity of Preferences. There is not a clear pattern of failure of quasi-concavity among the experiments. Although a fair number of responses are not neoclassical, their exclusion has little influence on the measures of central tendency.

6. Related Literature

Most of the studies that compare real and hypothetical responses have used closed-ended WTP questions, in which subjects could buy an item (or items) at a given price (Cummings *et al.*; Cummings, Harrison, and Rutström; Dickie, Fisher, and Gerking; see also the review in Foster, Bateman, and Harley). These studies can provide detailed information about the levels of WTP only when assumptions are made about the underlying value distribution or when both the sample size and the set of offer prices is large. Neither of those remedies is as appealing as an open-ended survey such as the BDM or second-price auction.

A few recent real open-ended studies were discussed in Section 4. The majority of real valuation experiments, however, have been studies of lotteries, which have been examined primarily for their implications about risk attitudes aversion rather than what they actually reveal about values. Much of the recent literature on choice under uncertainty has focused on pairwise choices (*e.g.*, Harless and Camerer; Hey and Orme) rather than on open-ended questions.

Most real experiments, including this paper's, are not fully real but are what might be called "randomly real." In random real experiments, subjects answer several questions, only one of which takes place for real money. In Shogren *et al.*'s candy bar experiments, one out of a subject's five bids took place for real; in their contaminated sandwich experiments, one out of twenty bids took place. In Bateman *et al.*'s Coke and chocolates experiments, one out of eighteen bids took place for real. In the experiments reported in this paper, either one out of two or one out of three of a subject's bids took place for real.

Karni and Safra show that the BDM mechanism is not everywhere incentive compatible in eliciting the certainty equivalents of non-degenerate lotteries when subjects have non-expected utility preferences. (See also Holt, who recognized the role of the independence axiom when only one out of out of a set of choice experiments is conducted for real, and subjects apply the reduction principle.) The BDM is incentive compatible, however, in eliciting true reservation values for nonrandom bundles, provided only that the subject prefers first stochastic dominance.

The BDM mechanism may, however, not elicit true preferences when subjects anticipate making errors in stating their values. If a subject anticipates making an error, the error introduces an element of risk into the revelation decision. She can reduce this risk by stating a higher mean compensation demanded (*i.e.*, before the error is drawn), since a higher CD reduces the probability that she will return her items and receive the random payment; that is, a higher CD exposes the subject to less risk. If the subject is risk averse *and* if she believes that the distribution of price offers has a mean equal to her true mean CD, then she would adopt the strategy of stating a CD that on average is higher than her true value. The size of the bias will be greater the more

risk averse is the individual and the higher is her error variance. Both elements should be small in our experiments. When errors in stated values are possible, the welfare implications of the "errors" themselves are probably much more important than their possible effect on incentive properties and observability.

7. Concluding Remarks

Although the values elicited by real valuation experiments are presumably the building-blocks of economics, those values have rarely themselves been the subject of analysis. Most studies have compared results across experimental treatments rather than focused on the actual outcomes of a given treatment. This study has looked at the responses that some real experiments elicit.

We have focused on the economic plausibility of those responses, which in our experiments depends on the resale values and replacement costs of the items. This study has been generous in its assessments of resale and replacement costs. A more realistic assessment of opportunity costs would result in many more responses being dropped. Even with the rather loose bounds we apply for economic plausibility, many subjects in our real experiments exceeded those bounds; that is, they ignore opportunity costs. We find subjects who appear to ignore either total or marginal opportunity cost. Since opportunity costs are essential to economic logic, their absence in experiments involving real money and real goods (of comparatively high value) casts doubt on the use of real experiments as the sole indicators of preferences or as a standard against which hypothetical surveys should be judged. Supporters of real experiments are left with two alternatives.

First, real experiments may be treated as inviolable – researchers can ask

positive questions but not apply normative judgments. In other words, the experimenter may be considered to have given up the right to reject any subject's response in a real experiment, because such a decision runs counter to the claim that these responses are true reflections of value and welfare.

Second, researchers can drop questionable observations and then be satisfied with valuations from only a portion of the population, perhaps as little as one-half, with the other responses dismissed as either irrational or "noisy" but inconsequential for policy analysis. This is analogous to being satisfied with only some experimental procedures; namely, single-item, closed-ended, willingness-to-pay experiments.

The preferred response, we believe, is to recognize that even the best and most realistic experiments will yield different pictures of "preferences." Economists should build connections between these different expressions of preference. Such an approach should allow both positive and normative exploration of economic behavior.

Table 1. Compensation Demanded for Binoculars

Observation #	CD for 2 Binoculars	CD for 3 Binoculars
1	$3	$5
2	5	7
3	5	10
4	10	10
5	10	15
6	12	18
7	13	20
8	15	15
9	15	22.50
10	16.25	25.50
11	18	27
12	19	39.50
13	20	30
14	20	35
15	20	40
16	20	30
17	20	35
18	20	30
19	20.25	30.50
20	22.50	35
21	25	50
22	25	20
23	29	45
24	30	45
25	30	45
26	35	55
27	35	50
28	39	70
29	40	60
30	40	60
31	40	60
32	40	60
33	42.50	60.75
34	45	67.50
35	50	75
36	50	75
37	60	90
38	80	105
39	100	100
40	100	100
41	100	150
42	175	325
43	200	200
44	250	370
45	300	450
46	1000	1500
47	10 million	20 million
48	20 million	60 million

Table 2. Compensation Demanded for Flashlights

Observation #	CD for 1 Flashlight	CD for 2 Flashlights	CD for 3 Flashlights
1	$1	$2	$1
2	1	1.25	1.50
3	2	4	9
4	2.25	4.50	6.75
5	3	6	N[a]
6	3	10	20
7	3	7	15
8	3	6	9
9	3	6	N
10	3	8	12
11	3.50	6	10
12	3.50	3.25	4.50
13	4	8	12
14	4.25	8.50	12.75
15	4.75	7.50	10
16	5	9.50	14
17	5	8	10
18	5	7	8
19	5	10	30
20	5	9.50	10
21	5	7.50	10
22	5	5	5
23	5	10	15
24	6	12	16
25	6	12	15
26	6	7	8
27	6	12	20
28	6	9	12
29	7	12	18
30	7.50	10	15
31	8.25	16.50	24.75
32	8.50	17	30
33	9.50	14.75	26
34	10	20	N
35	10	20	30
36	12	25	40
37	12	10	18
38	12.50	15	45
39	16	26.95	38
40	19.95	38.95	40
41	N	N	N
42	N	N	N

[a]Subjects were instructed as follows: "If you don't want to sell your flashlights back at any price (you definitely want to keep all 3), just put an N in the blank."

Table 3. Compensation Demanded for Mugs

Observation #	CD for 2 Mugs	CD for 3 Mugs
1	$2	$2.50
2	2	3
3	2	3
4	3	5
5	3	5
6	3.50	5
7	4	6
8	5	7.50
9	5	7.50
10	6	9
11	6	9
12	6	9
13	6	9
14	6	15
15	6.50	10
16	7.50	12
17	7.98	12.97
18	8	10
19	8	10
20	8	15.75
21	8.50	10
22	9	13.50
23	9.95	14
24	10	14.25
25	10	15
26	10	15
27	10	15
28	10	15
29	10	15
30	10	15
31	12	18
32	12.50	18.50
33	15	18
34	15	20
35	15	21.50
36	15	22
37	15	25
38	16	22
39	17.99	25
40	20	20
41	20	28.50

References

Bateman, I., A. Munro, B. Rhodes, C. Starmer, and R. Sugden, "A Test of the Theory of Reference-Dependent Preferences," *Quarterly J. Econ.* 112 (1997) 479-505.

Cummings, R., S. Elliott, G. Harrison, and J. Murphy, "Are Hypothetical Referenda Incentive Compatible?" *Journal of Political Economy* 105 (1997) 609-21.

Cummings, R., G. Harrison, and E. Rutström, "Homegrown Values and Hypothetical Surveys: Is the Dichotomous-Choice Approach Incentive-Compatible?" *American Economic Review* 85 (1995) 260-66.

Dickie, M., A. Fisher, and S. Gerking, "Market Transactions and Hypothetical Demand Data: A Comparative Study," *J. Amer. Stat. Assoc.* 82 (1987) 69-75.

Foster, V., I. Bateman, and D. Harley, "Real and Hypothetical Willingness to Pay for Environmental Preservation: A Non-Experimental Comparison," *Journal of Agricultural Economics* 48 (1997) 123-38.

Haab, T., J.-C. Huang, and J. Whitehead, "Are Hypothetical Referenda Incentive Compatible? A Comment," *Journal of Political Economy* (forthcoming).

Haab, T., and K.E. McConnell, "Referendum Models and Negative Willingness to Pay: Alternative Solutions," *J. Environ. Econ. & Mgmt.* 32 (1997) 251-70.

Harless, D., and C. Camerer, "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica* 62 (1994) 1251-90.

Hey, J., and C. Orme, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica* 62 (1994) 1291-1326.

Holt, C., "Preference Reversals and the Independence Axiom," *American Economic Review* 76 (1986) 508-15.

Horowitz, J.K., and K.E. McConnell, "A Review of WTA/WTP Studies," Unpublished manuscript (1998).

Horowitz, J.K., K.E. McConnell, and J. Quiggin, "A Test of Competing Explanations of Compensation Demanded," *Economic Inquiry* (forthcoming).

Johannesson, M., B. Liljas, and R. O'Conor, "Hypothetical versus Real Willingness to Pay: Some Experimental Results," *Applied Economics Letters* 4 (1997) 149-51.

Karni, E., and Z. Safra, "Preference Reversal and the Observability of Preferences by Experimental Methods," *Econometrica* 55 (1987) 675-85.

Loomis, J., T. Brown, B. Lucero, and G. Peterson, "Improving Validity Experiments of Contingent Valuation Methods: Results of Efforts to Reduce the Disparity of Hypothetical and Actual Willingness to Pay," *Land Economics* 72 (1996) 450-61.

Neill, H., R. Cummings, P. Ganderton, G. Harrison, and T. McGuckin, "Hypothetical Surveys and Real Economic Commitments," *Land Economics* 70 (1994) 145-54.

Shogren, J., S. Shin, D. Hayes, and J. Kliebenstein, "Resolving Differences in Willingness to Pay and Willingness to Accept," *Amer. Econ. Rev.* 84 (1994) 255-70.