USING *EX ANTE* APPROACHES TO OBTAIN CREDIBLE SIGNALS FOR VALUE IN CONTINGENT MARKETS: EVIDENCE FROM THE FIELD

CRAIG E. LANDRY AND JOHN A. LIST

While contingent valuation remains the only option available for measurement of total economic value of nonmarketed goods, the method has been criticized due to its hypothetical nature. We analyze field experimental data to evaluate two *ex ante* approaches to attenuating hypothetical bias, directly comparing value statements across four distinct referenda: hypothetical, "cheap talk," "consequential," and real. Our empirical evidence suggests two major findings: hypothetical responses are significantly different from real responses; and responses in the consequential and cheap talk treatments are statistically indistinguishable from real responses. We review the potential for each method to produce reliable results in the field.

Key words: hypothetical bias, referendum, stated preference.

DOI: 10.1111/j.1467-8276.2007.01017.x

Accurately estimating the economic value of nonmarketed goods and services is essential for efficient public policy. While markets routinely provide signals of value for traded commodities, estimating values for goods and services that are not traded in markets provides a quandary for the policymaker. On the one hand, she can make use of market signals to estimate use values by utilizing revealed preference methods, such as travel cost or the hedonic approach. Alternatively, she may take a more holistic view and use a stated preference approach (e.g., contingent valuation), which to date is the only method that is capable of measuring the total economic value (use and nonuse) of a nonmarketed commodity. Yet, this approach presents its own set of challenges. In particular, some commentators have argued that contingent surveys are unreliable due to their hypothetical nature (e.g., Diamond and Hausman 1994).

Following Bohm's (1972) seminal work on estimating the demand for public goods, several dozen experimental studies have been undertaken to elucidate the relationship between hypothetical and real statements (see the literature review in List and Shogren [2002] and Harrison and Rutström [2005]). The weight of the evidence in this body of literature suggests that hypothetical bias—a divergence between behavior in real and hypothetical institutions—is often present, the implication being that it could be a significant problem for stated preference methods that use contingent markets. In response, economists have searched for ways to attenuate this bias. Following the recommendations of NOAA panel on contingent valuation (Arrow et al. 1993), Loomis, Gonzalez-Caban, and Gregory (1994) attempted to mitigate hypothetical bias by reminding respondents of their budget constraint and highlighting substitutes. While the authors find no evidence that these subtle changes in the survey instrument have an effect on subject responses, budget constraint and substitute commodity reminders have become standard practice for stated preference methods.

Other forms of *ex ante* adjustment of survey instruments were subsequently explored.[1] Cummings, Harrison, and Taylor (1995) introduced what has come to be known as "cheap

The authors are, respectively, Assistant Professor, Department of Economics, East Carolina University and Professor, Department of Economics, University of Chicago and NBER. Thanks to the Editor and three anonymous reviewers who provided comments that improved the paper. Glenn Harrison, Jason Shogren, and Laura Taylor also provided comments throughout the research process.

[1] At the same time, economists have been exploring *ex post* alternatives to addressing hypothetical bias, which involve statistical calibration of responses. See Blackburn, Harrison, and Rutström (1994), Champ et al. (1997), Fox et al. (2003), and List and Shogren (2002). Results generally suggest that calibration factors are commodity-specific. Thus, calibration may not be flexible enough to provide a general approach to attenuating hypothetical bias.

Amer. J. Agr. Econ. 89(2) (May 2007): 420–429

talk" in the nonmarket valuation literature.[2] "Cheap talk" is a script that is instituted before value elicitation. The "cheap talk" script: (i) describes hypothetical bias and provides an example; (ii) reviews possible explanations for such bias; and (iii) encourages subjects to vote as if the valuation question were real (i.e., had real economic consequences). Cummings and Taylor (1999) test the "cheap talk" script against real and hypothetical referenda for contributions to public goods. They find strong evidence in support of this approach. In their trials where hypothetical bias was found, the bias largely disappeared when the "cheap talk" script preceded the hypothetical valuation question; voting behavior was not statistically different across real referenda and hypothetical referenda that included the cheap talk script.

The success of cheap talk has not been universal, however. List (2001) and Lusk (2003) find that a cheap talk script is effective in attenuating hypothetical bias only for certain classes of subjects—those with less market experience or less familiarity with the good being valued. Aadland and Caplan (2003) have been successful in attenuating hypothetical bias with a shortened cheap talk script, whereas previous research found that a shortened cheap talk script was not effective (Cummings, Harrison, and Taylor 1995; Poe, Clark, and Schulze 1997).

Other *ex ante* methods that have been introduced to attenuate hypothetical bias include a learning model in which respondents gain experience with the valuation mechanism in a real setting before the hypothetical setting is introduced. Bjornstad, Cummings, and Osbourne (1997) find that participation in a real referendum preceding a hypothetical one induces behavior in the subsequent hypothetical setting that is not distinguishable from behavior in a real referendum. Smith and Mansfield (1998) find similar results with a dichotomous choice mechanism.

While this lot of studies certainly has value, in contingent valuation surveys carried out in the field it is commonplace to present respondents with a realistic scenario, inducing them *to believe* that their responses have a degree of importance associated with them. Accordingly, in contingent markets it appears reasonable to assume that individuals' beliefs about whether their responses will actually be considered in policy circles varies—some may believe with a high degree of certainty that their responses are important, whereas others may have significant doubts. This characterization stands in stark contrast to that found in the studies cited above (i.e., real versus hypothetical statements of value).

This idea, identified as "realism" by Cummings and Taylor (1998) and "consequentialism" by Carson, Groves, and Machina (2000), suggests that stated preference survey designs that are "realistic" will induce subjects to truthfully reveal their preferences.[3] As discussed by Carson, Groves, and Machina, a binary choice referendum will be incentive-compatible assuming: (i) a weakly monotonic influence function (i.e., a higher proportion of supporting votes will not decrease the probability of provision), (ii) a coercive payment mechanism, and (iii) a closed valuation mechanism (i.e., the good cannot be provided in another way). The intuition is that if subjects believe that their responses have the potential to influence public policy, then there is no incentive for them to misrepresent their preferences. The "consequential" design approach can be applied in a straightforward manner: inform subjects that their responses matter in a probabilistic sense and they should truthfully reveal their preferences.[4]

The only papers of which we are aware that explore such a mechanism in an experimental setting are Cummings and Taylor (1998) and Carson et al. (2002), both of which utilize a referendum format. The results of Cummings and Taylor suggest that treatments utilizing low levels of probability ($p \leq 0.5$) to link voting behavior to real economic consequences produce results not in accord with a binding referendum ($p = 1.0$), but voting behavior associated with higher probability levels ($p = 0.75$) cannot be distinguished from that of a binding referendum. On the other hand, Carson et al. find that subjects voting in probabilistic referenda

[2] Loomis et al. (1996) utilize a similar approach in their experiments on hypothetical bias with private goods that are readily available in the market place. They appeal to subjects not to provide an estimate of the market price of the good in their value elicitation experiments. They find that such appeals do attenuate hypothetical bias somewhat.

[3] We stick with the "consequentialism" moniker to distinguish this treatment from real treatments.

[4] This does not suggest outright deception. Rather, if the findings may influence public policy, then this should be relayed to the respondents. Note the similarities between this methodology and the "randomized payment" approach used in experimental economics, whereby agents play, for example, ten rounds of a game and are only paid for one round, which is determined randomly.

(where probabilities of the referendum binding range from $p = 0.20$ to $p = 0.80$) do not behave differently than subjects voting in a binding referendum ($p = 1.0$).

While both cheap talk and consequentialism appear to have enjoyed a degree of success in gathering economic values that correspond to values obtained in binding elicitation mechanisms, to our knowledge no study has systematically compared responses across these *ex ante* methods with an incentive compatible instrument. What we offer in this article is precisely such a comparison.[5] To provide insights into the effectiveness of these *ex ante* methods within an otherwise identical protocol that is incentive compatible, we make use of a straightforward 2×4 experimental design with 256 subjects from an real marketplace—the sports card market. In order to foster incentive-compatibility, we incorporate the experimental design of Carson et al. (2002), which uses a majority voting mechanism that determines the transfer of n payments of a prespecified amount of money from the subjects to the experimental monitor, coupled with the delivery of n private goods to the subjects. The transfers of n pieces of sports memorabilia simulates the provision of a public good in the sense that either all n subjects pay the prespecified amount and receive an identical piece of sports memorabilia or none do. A coercive payment mechanism is utilized, and use of a private good ensures that the referendum is incentive compatible. Using identical written protocol, we conduct four distinct referenda: hypothetical, hypothetical-with-cheap-talk, consequential, and real.

Comparing behavior across these four treatments, we report two major findings. First, consistent with many other experimental results, our experimental evidence suggests that responses in the hypothetical referenda are significantly different from responses in the real referenda. Second, responses in the consequential and hypothetical with cheap talk treatments are, for the most part, statistically indistinguishable from responses in the real referenda. Yet, the data do hint that responses in the hypothetical with cheap talk treatments represent an upper bound on real responses. Our tentative conclusion is that accurate signals of value are most likely obtained from the subjects that view their decisions as being sufficiently consequential. However, since in the field the perception of consequences is subjective, the cheap talk design is likely to be a useful alternative, especially in those cases where the likelihood of successfully achieving consequentialism is small.

The remainder of this article proceeds as follows: Section 2 summarizes the experimental design; Section 3 discusses the results; Section 4 concludes with a discussion of how these results can potentially aid public policy decision making.

Experimental Design

Our field experiment was conducted on the floor of a sports memorabilia show in Tucson, Arizona. As discussed in previous work (e.g., List 2001), with the rise in popularity of sports cards and memorabilia in the past two decades, markets have naturally arisen that allow for the interaction of buyers and sellers. The physical marketplace is typically a gymnasium or hotel conference center. When the market opens, consumers mill around the marketplace bargaining with dealers, who have their merchandise prominently displayed. The duration of a typical sports card show is a weekend, and a lucrative show may provide any given dealer hundreds of exchange opportunities (buying, selling, and trading of goods).

On the weekend in which we ran our field experiment, we approached attendees as they entered the sports card show and inquired about their interest in participating in an experiment. The interceptor explained to each potential subject that they would receive \$10 for showing up if they decided to participate. Upon obtaining an agreement to participate, the interceptor informed the subject of the time and place of the experiment (a reserved room in the hotel conference center). Each subject was allocated to one, and only one, of the eight sessions. (As described below, each of the eight sessions represented a distinct treatment.)

Upon arrival to the experimental session, individuals signed a consent form upon which they agreed to abide by the rules of the experiment, received their \$10 show-up payment,

[5] We note that two separate papers by Cummings and Taylor (1998, 1999) test "realism" and "cheap talk" in the same institution with precisely the same good, but straightforward comparisons of the methods have not been highlighted in the literature. Also, as noted by Taylor (1998), the referendum used in these papers is not closed, and therefore not incentive compatible. Bulte et al. (2005) explore cheap talk and consequentialism within the same experimental design but unfortunately have no actual values due to the nature of their good.

Table 1. Experimental Design—Subjects by Treatment and Price Sequence

Treatment	Hypothetical	Cheap Talk	Consequential	Real
A: \$5/\$10	30	32	29	33
B: \$10/\$5	34	37	30	31

and were given experiment instructions. Depending on the session in which they participated, they were allocated randomly to one of the eight treatments summarized in table 1. Table 1 presents a summary of our 2×4 experimental design and provides sample sizes in each treatment. Table 1 can be read as follows: columns indicate treatment type—hypothetical, hypothetical-with-cheap-talk, consequential, or real; rows indicate pricing sequence—\$5/\$10 (pricing sequence *A*), or \$10/\$5 (pricing sequence *B*).

Taking one of these treatments as an example, consider an excerpt from the instructions of the real treatment (full instructions are available upon request):

> Welcome to Lister's Referendum. Today you have the opportunity to vote on whether 'Mr. Twister,' this small metal box, will be 'funded.' If 'Mr. Twister' is funded, I will turn the handle and *n* (the amount of people in the room) ticket stubs dated October 12, 1997, which were issued for the game in which Barry Sanders passed Jim Brown for the number 2 spot in the NFL all-time rushing yardage, will be distributed—one to each participant (illustrate). To fund 'Mr. Twister,' **all** of you will have to pay \$*X*.

We utilized a referendum vote for the provision of a public good as our value elicitation mechanism, as this is a common method utilized in field applications of stated preference methods and closed-ended mechanisms were recommended by the NOAA panel on contingent valuation (Arrow et al. 1993). As implied above, we obtain voting responses from each subject for *each* of two price levels (i.e., \$*X* was \$5 in one question and \$10 in the other). For example, subjects in the real treatment pricing sequence *A* first provide a response for the \$5 question and then for the \$10 question, regardless of how the group responded to the initial \$5 offer.[6] By varying the price level in such a manner, our data allow for within and between comparisons of the effect of offer price on voting behavior, in addition to between comparisons of treatment effects. We change the ordering of the offer prices to test for sequencing effects, as exhibited in the rows of table 1.

Our referendum mechanism operates on a simple majority vote for *n* identical pieces of sports memorabilia (where *n* equals the number of subjects). In the real treatment, these *n* private goods are provided to everyone in the experiment if a majority of the subjects vote to "fund" the public good "Mr. Twister," while provision is made in the consequential treatment if the referendum vote is binding (determined by random outcome from known probability distribution). In this way, the *n* private goods simulate the provision of a public good—no one is excluded from the provision of the sports memorabilia (or lack thereof), and the consumption of the *n* items is nonrival because there are precisely enough items to go around—one for each subject. In order to avoid free-riding and focus attention on hypothetical bias, we use a coercive payment mechanism and a private good.

In the hypothetical treatments, following previous efforts, we used passive language so subjects understood that their vote would not induce true economic consequences—i.e., no money or goods would change hands. The cheap talk treatments were also hypothetical, but included a "cheap talk" script (as described above). The language in the cheap talk script is originally from Cummings and Taylor (1999),

[6] Carson, Groves, and Machina (2000) suggest that the use of two or more prices in value elicitation could (i) imply uncertainty of price, (ii) imply a willingness to bargain on behalf of the seller, or (iii) induce a perceived change in quantity/quality of the good. Price uncertainty would decrease the median or mean valuation (from the second question) for risk-averse agents, while the direction of change for the latter cases depends upon the response to the initial question. In the discussion of Carson, Groves, and Machina, however, the magnitude of the second price is always conditional on the initial response. Those who reply "No" are offered a lower price, while those who indicate "Yes" are asked a higher price. In all of these cases, the introduction of a second price signals that something else could be going on—the transaction involves more than is apparent at face value. Our price sequences, in contrast, are a design parameter that is purely exogenous. The sequence of prices is not conditional upon the response of the subjects, and the prices are offered aloud, for all to hear. Moreover, in cases of a potential real payment, it is made clear that the binding price will be determined randomly. These attributes of our experiment could attenuate any strategic responses of our subjects.

with necessary changes due to differences in the allocation mechanism and good. In the consequential treatments, subjects were told that separate coin flips would determine: (i) which of the price levels ($5 or $10) would be utilized, and (ii) whether the corresponding votes would be economically binding. Hence, once the price level was chosen, the probability that the subjects' voting responses had real consequences was 50%. The real treatment was a straightforward referendum, but, again, since the agent was voting on the same good twice (for both $5 and $10), a coin-flip was used to determine which price level was binding, after the subjects had indicated their vote at *both* price levels.

A few noteworthy items should be mentioned before we proceed to the experimental results. First, all of our subjects were "ordinary" consumers (i.e., none of the experimental subjects were sports card dealers). Second, as aforementioned, subjects participated in only one treatment. Third, the experimenter was careful not to examine the votes from the first price level before asking subjects to vote in the second referendum at the second price level.

Results

A summary of the experimental data is provided in table 2. Our first order of business is to examine the field data for internal consistency. This is important since a recent study (Ariely, Loewenstein, and Prelec 2003—ALP hereafter) suggests that valuation experiments can produce results that appear coherent in the sense that subjects are responsive to within-session variation in quantities or prices, but are arbitrary in that the valuations are conditioned on design parameters that should be irrelevant to fundamental values. In one of their experiments, ALP find behavior consistent with downward-sloping demand curves when examining data associated with the same individual (a within comparison), but they found that the expressed value for common consumer products, measured with a theoretically incentive-compatible mechanism, can be considerably influenced by exposure of the subject to a clearly uninformative, random anchor.

To test for internal consistency within treatments, we utilize the binomial test (the small sample analog of McNemar's exact test for the equality of correlated proportions), the null hypothesis being that the proportion of "Yes" responses at $5 is equivalent to the proportion of "Yes" responses at $10. The alternative hypothesis is that the proportion of "Yes" responses is larger at the $5 level. At the $p < 0.10$ level, we reject the null hypothesis for both pricing sequences in the cheap talk and real treatments.[7] We, likewise, reject this hypothesis for pricing sequence A in the consequential treatment, but not pricing sequence B.[8] In the latter case, we find no evidence of behavior inconsistent with demand theory (i.e., a "Yes" response to $10 followed by a "No" response to $5), rather there was little response to the price change (only two subjects changed from "No" to "Yes" as the price fell from $10 to $5). We fail to reject the hypothesis of equal proportions of "Yes" responses in both pricing sequences for the hypothetical treatment.[9] In the hypothetical treatments, there were three cases where subjects exhibited behavior inconsistent with demand theory.[10] If we ignore these responses, we reject the null hypothesis. Thus, our within-subject data are generally consonant with ALP and demand theory.

We test for between-subject demand consistency by examining equality of proportions at different prices across the pricing sequences using a chi-square test. That is, we compare responses to the $5 question in sequence A to responses to the $10 question in sequence B, and vice versa, for each of the treatments. In only two cases (comparing cheap talk A:$10 and B:$5, and comparing real A:$10 and B:$5)[11] can we reject the hypothesis that these responses are equivalent. In the six other cases we fail to reject this hypothesis.[12] Visual inspection of the data confirms

---

[7] Cheap talk treatment, pricing sequence A p-value = 0.0313; cheap talk treatment, pricing sequence B p-value = 0.0078; real treatment, pricing sequence A p-value = 0.0313; real treatment, pricing sequence B p-value = 0.0625.

[8] Consequential treatment, pricing sequence A p-value = 0.0313; consequential treatment, pricing sequence B p-value = 0.25.

[9] Hypothetical treatment, pricing sequence A p-value = 0.1563; hypothetical treatment, pricing sequence B p-value = 0.1641.

[10] Our results are similar to those of Lusk (2003), in which responses derived from an elicitation mechanism that utilized cheap talk exhibited more responsiveness to price than those without cheap talk (i.e., hypothetical data). We find that both *ex ante* methods exhibited more responsiveness to variation in price than hypothetical data.

[11] χ^2 values are 3.9190 (p-value = 0.0477) and 6.3441 (p-value = 0.0118), respectively.

[12] Consider first other comparisons of A:$10, B:$5. χ^2 values are 1.7911 (p-value = 0.1808) for hypothetical treatment and 1.3371 (p-value = 0.2475) for consequential treatment. Consider next comparisons of A:$5, B:$10. χ^2 values are 0.4637 (p-value = 0.4959) for hypothetical, 2.1588 (p-value = 0.1455) for cheap talk, 0.9433 (p-value = 0.3314) for consequential, and 0.1383 (p-value = 0.7100) for real.

Table 2. Voting Behavior by Treatment

Treatment	Hypothetical		Cheap Talk		Consequential		Real	
Pricing sequence	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*
First offer price	\$5	\$10	\$5	\$10	\$5	\$10	\$5	\$10
Second offer price	\$10	\$5	\$10	\$5	\$10	\$5	\$10	\$5
Subjects (*n*)	30	34	32	37	29	30	33	31
	25	26	15	10	10	7	9	8
First	(0.83)	(0.76)	(0.47)	(0.27)	(0.34)	(0.23)	(0.27)	(0.26)
Yes	22	29	10	17	5	9	4	12
Second	(0.73)	(0.85)	(0.31)	(0.46)	(0.17)	(0.30)	(0.12)	(0.39)
Pooled *n*	64		69		59		64	
Pooled Yes \$5	54		32		19		21	
	(0.84)		(0.46)		(0.32)		(0.33)	
Pooled yes \$10	48		20		12		12	
	(0.75)		(0.29)		(0.20)		(0.19)	

Note: Proportions are indicated in parentheses.

that the proportions of "Yes" responses associated with the \$5 price level in one sequence are all higher than the proportions of "Yes" responses associated with the \$10 price level from the other pricing sequence. Nonetheless, in six out of eight cases this difference is not statistically significant. We infer that demand for our PSA graded mint ticket stubs is inelastic within the price range we offered.[13]

Next, we examine the proportion of "Yes" votes for each price level, across pricing sequences *A* and *B*. We use the χ^2 statistic to compare identical prices when offered first and second, for each of the four treatments. In contrast to the results of ALP, we find no evidence of significant sequencing effects in the voting proportions for either price level, suggesting that anchoring may not be an important phenomenon in the marketplace.[14] Thus, for efficiency purposes, we pool the data within price cells. The subsequent results use the pooled data, which are summarized in table 2.[15]

Treatment Effects

Figure 1 (Figure 2) provides a graphical depiction of the pooled data contained in table 2 by presenting the proportion of "Yes" votes across treatments for the \$5 (\$10) price level. The data paint an interesting picture: 32.8% of subjects voted to fund the public good in the real \$5 treatment, while in the \$5 consequential treatment, 32.2% voted "Yes." These proportions are notably similar. On the other hand, the proportion of "Yes" votes in the \$5 cheap talk treatment was considerably greater, at 46.4%, and the \$5 hypothetical treatment exhibited a much larger proportion of "Yes" votes: 84.4%. Similar trends are evident in the \$10 data.[16] Overall, perusal of table 2 and figures 1 and 2 suggests that voting behavior across the four referenda is considerably different. This insight is documented statistically, as the χ^2 (df = 3) statistic for the test for equality of the four proportions allows us to reject the homogeneity null at the $p < 0.01$ level for

[13] An anonymous reviewer points out that an implication of the results of ALP is that demand for price changes derived within subjects will be more elastic than that derived between subjects. Visual inspection of the data and the statistical results, in general, does not lend support to this hypothesis, but this clearly depends on the parameters and experimental design.

[14] χ^2 values for the \$5 price level are 0.04 ($p = 0.82$), 0.01 ($p = 0.93$), 0.13 ($p = 0.71$), and 0.94 ($p = 0.33$) for the hypothetical, cheap talk, consequential and real treatments, respectively (all df = 1). Corresponding χ^2 statistics for the \$10 price level are 0.08 ($p = 0.77$), 0.14 ($p = 0.69$), 0.33 ($p = 0.56$), and 1.9 ($p = 0.16$) (all df = 1).

[15] We also conducted all analyses using only the first responses (i.e., the \$5 responses from pricing sequence *A* and the \$10 responses from pricing sequence *B*); our primary conclusions do not change.

[16] Percentage of "Yes" votes in the \$10 real treatment = 18.8%; percentage of "Yes" votes in the \$10 consequential treatment = 20.3%; percentage of "Yes" votes in the \$10 cheap talk treatment = 29%; percentage of "Yes" votes in the \$10 hypothetical treatment = 75%.

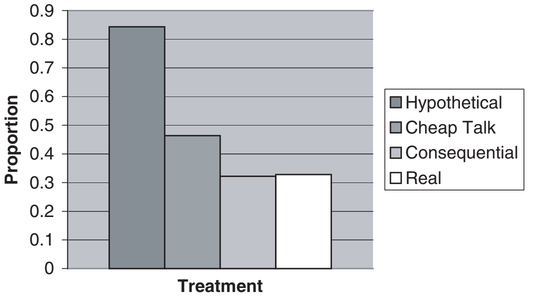

Figure 1. Proportion of "Yes" votes by treatment: $5 price level

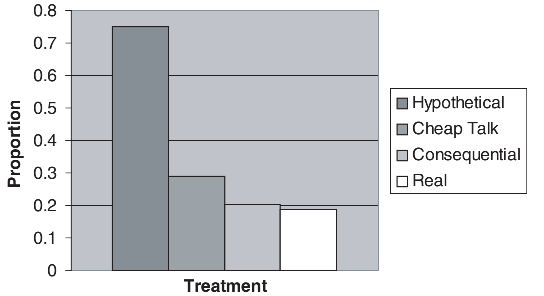

Figure 2. Proportion of "Yes" votes by treatment: $10 price level

both price levels ($5: $\chi^2 = 45.5980$ and $10: $\chi^2 = 58.3138$).

Turning to a comparison of the individual treatment effects, we present table 3, which summarizes statistics of pair-wise χ^2 tests. The upper right (lower left) triangular elements present the statistics for the $5 ($10) price level. A first important question is whether voting behavior in the hypothetical treatments is different from voting patterns in the real treatments. The raw data suggest large differences between the hypothetical referendum and the actual referendum: whereas 84% (75%) voted "Yes" to the proposition in the hypothetical treatments at the $5 ($10) offer price, only 33% (19%) voted "Yes" in the real treatment. Indeed, as is presented in table 3, the proportion of affirmative votes in the hypothetical referendum is statistically different from the percentage of affirmative votes in the real treatment (as well as the other two treatments) at the $p < 0.01$ level.[17] Thus, our evidence suggests that subjects' respond differently in hypothetical referenda than they respond in our three other types of referenda.

Turning to comparisons of data from other treatments, we find that voters in the cheap talk treatment tend to vote "Yes" more often in both the $5 and $10 treatments compared to voters in the real treatment. While this pattern is stark, these observed differences are not statistically significant at conventional levels: for the $5 offer price, the χ^2 (df = 1) statistic for real versus cheap talk is 2.5486 (p-value = 0.1104), and for the $10 price level, the χ^2 (df = 1) statistic for real versus cheap talk is 1.9038 (p-value = 0.1677). Yet, it should be noted that using a one-sided alternative, these differences are significant at the $p < 0.10$ level.

Since valuation experiments are typically utilized to estimate willingness to pay (WTP) we are interested in whether data from these different treatments produce comparable measures in this regard. We used the nonparametric Turnbull to estimate the lower bound of mean WTP (Haab and McConnell 2002). In doing so, we utilized data on the response to only the first price offered in each price sequence (since the Turnbull requires independence of responses to randomly assigned prices). We find that we cannot reject the null hypothesis $WTP_{\text{cheap talk}} = WTP_{\text{real}}$ at the $p =$ 0.1921 level ($t = 1.3091$, df = 131) for a two-tailed test. But, again, if we use a one-sided alternative, we reject the equality of lower bound WTP estimates at $p < 0.10$.

We find that affirmative responses in the consequential and real treatments are roughly equivalent: 32.2% (20.3%) in the $5 ($10) offer price versus 32.8% (18.8%), respectively. For the $5 offer price, the χ^2 (df = 1) statistic for real versus consequential is 0.2594 (p-value = 0.6105). For the $10 offer price, the χ^2 (df = 1) statistic for real versus consequential is 0.04934 (p-value = 0.8215). We therefore cannot reject the hypothesis that these data are the same at conventional significance levels. In addition, we cannot reject the null hypothesis that the Turnbull estimates of the lower bound of mean WTP are equal ($t = 0.2941$, df = 121; p-value = 0.7692). The evidence is in favor of the consequential design's ability to provide reliable signals of value.

As a final test, we compare data from the cheap talk and consequential treatments. For the $5 offer price, the χ^2 (df = 1) is 2.6656

[17] χ^2 (df = 1) statistics for the hypothetical versus cheap talk, consequential, and real treatments are 20.9802, 34.6348, and 35.0672, respectively, for the $5 price level. For the $10 price level, the χ^2 (df = 1) statistics are 28.1351, 36.7114, and 40.6588 for the same sequence of tests. All p-values are below 0.0001.

Table 3. Experimental Statistics for Pair-Wise Comparisons across Treatments

Treatment	Hypothetical	Cheap Talk	Consequential	Real	
Hypothetical	–	20.9802 (0.0000)	34.6348 (0.0000)	35.0672 (0.0000)	\$5 Price level
Cheap talk	28.1351 (0.0000)	–	2.6656 (0.1025)	2.5486 (0.1104)	\$5 Price level
Consequential	36.7114 (0.0000)	1.2682 (0.2601)	–	0.2594 (0.6105)	\$5 Price level
Real	40.6588 (0.0000)	1.9038 (0.1677)	0.04934 (0.8215)	–	
	\$10 Price level	\$10 Price level	\$10 Price level		

Note: The upper triangle contains test statistics for the \$5 price level; the lower triangle contains test statistics for the \$10 price level. All df = 1, and p-values are in parentheses.

(p-value = 0.1025), and for the \$10 price it is 1.2682 (p-value = 0.2601). We therefore cannot reject the null hypothesis that the data are derived from the same underlying parent population for either offer price at conventional significance levels. Moreover, turning to the Turnbull estimate mean WTP, we find evidence that suggests we should not reject the null hypothesis $WTP_{\text{cheap talk}} = WTP_{\text{consequential}}$ ($t = 0.9814$, df = 122; p-value = 0.3283).[18]

Discussion and Conclusions

Whether contingent markets can produce credible value estimates remains of utmost policy importance. Indeed, for public regulators and damage assessors, contingent surveys remain the only method that can potentially obtain estimates of total economic value for nonmarketed commodities. Using data gathered from more than 250 subjects, we find experimental evidence that suggests responses in hypothetical referenda are significantly different from responses in real referenda. This result is in accordance with many of the studies that have examined hypothetical and real statements of value. Yet, we do find evidence that when decisions *potentially* have financial consequences, subjects behave in a fashion that is consistent with behavior when they have consequences with certainty. Our results furthermore suggest that estimates of the lower bound of mean WTP derived from "consequential" referenda are statistically indistinguishable from estimates of the actual lower bound of WTP.[19]

Such insights represent good news for stated preference surveys, as a necessary condition for their efficiency is that they are able to provide accurate estimates of value. Yet, this news should be tempered in that such results represent only the beginning of the research process. Even if our results are found to hold across different experimental designs and other types of manipulations the necessary next step is ensuring that survey respondents view the instrument as consequential. In our experiment and other related laboratory exercises (i.e., Cummings and Taylor 1998), the probabilities utilized are clearly objective, being defined by the experimental monitor in a transparent way (the appropriate mix of different colored bingo balls or specific outcomes associated with the roll of a 10-sided die or coin flip). In the field, beliefs about a contingent referendum vote actually affecting policy are subjective, largely out of the control of researchers.

Utilizing postsurvey questionnaires, previous research suggests that survey respondents' believe that the money generated would actually be spent on the proposed project (Powe, Garrod, and McMahon 2005) and that the majority of respondents regard the CV results as something that is likely to be of use to policy makers (Brouwer et al. 1999). However,

[18] Using an F-test, we cannot reject the hypotheses $Var(WTP_{\text{consequential}}) = Var(WTP_{\text{real}})$ (F = 1.1285, p-value = 0.3186), $Var(WTP_{\text{hypothetical}}) = Var(WTP_{\text{real}})$ (F = 1.2283, p-value = 0.2084), or $Var(WTP_{\text{cheaptalk}}) = Var(WTP_{\text{real}})$ (F = 1.0759, p-value = 0.3853).

[19] A natural question concerning our consequentialism results is why they are different from Cummings and Taylor (1998), who report that treatments utilizing low levels of probability ($p \leq 0.5$) produce results not in accord with a binding referendum ($p = 1.0$), but voting behavior associated with higher probability levels ($p = 0.75$) cannot be distinguished from that of a binding referendum. This remains an open empirical question, as Harrison and List (2004) point out when making a similar comparison to motivate the use of field experiments and what might cause differences between the lab and the field: "To provide a direct example of the type of problem that motivated us, when List [2001] obtains results in a field experiment that differ from the counterpart lab experiments of Cummings, Harrison, and Osborne [1995] and Cummings and Taylor [1999], what explains the difference? Is it the use of data from a particular market whose participants have selected into the market instead of student subjects, the use of subjects with experience in related tasks, the use of private sports-cards as the underlying commodity instead of an environmental public good, the use of streamlined instructions, the less intrusive experimental methods, mundane experimenter effects, or is it some combination of these and similar differences?"

we are unaware of any results in the stated preference literature that offer an explicit assessment of perceived consequences of survey respondents. Thus, it strikes us that another important focus of future research should be to assess perceived consequences of survey respondents subsequent to value elicitation and learn about the factors that influence such perceptions. While we are unaware of how various procedures increase the likelihood of consequentialism, stated preference researchers generally realize the importance of providing background information on the public good of interest and policy options available for addressing its provision, which might heighten consequentialism. We cannot emphasis enough the importance of pretesting surveys in order to improve the perception of realism on the part of respondents.

In addition, practitioners of stated preference should continue to focus on the realism associated with payment vehicles (the hypothetical method by which payment for the public good would be made). For example, higher overall price levels may not seem tied to public good provision in a realistic way, but on the other hand, higher electricity prices, taxes, or the institution of user fees probably will. As suggested above, debriefing questions can help to improve the understanding of respondents' perceptions of the survey questions. A simple Likert-scale assessment of perceived consequences (i.e., level of agreement/disagreement with some statement regarding the likelihood that survey responses will influence the eventual policy decision) could be quite informative and not likely to be onerous or costly to collect.

Given the potential problems in designing "consequential" stated preference surveys, we also highlight our results regarding the effectiveness of the "cheap talk" design. Our experimental evidence does support the cheap talk design, but it does not appear as strong as the consequential design (with an objective probability of $p = 0.50$). However, in actual applications of stated preference methods cheap talk provides an important alternative to the consequential design in cases where realism is difficult to attain, or in cases where the variability in perceptions of realism tend to be high. We note that such conditions could be quite common in the field, and thus cheap talk remains a viable design option.

Important extensions of this research include implementing the consequential design with different probability levels, making allowances for subjective or uncertain probabilities, and incorporating goods with a nonuse component. Our field data make use of subjects that are familiar with the class of good being valued (presumed since they have self-selected into the market for sports memorabilia), and arguably the good conveys primarily use value. Since part of the value of stated preference surveys stems from their purported ability to measure nonuse value, it is of interest to know whether referenda for potentially unfamiliar goods with primarily nonuse value will produce comparable results to those of this paper. This is a topic for future research.

*[Received May 2005;
accepted May 2006.]*

References

Aadland, D., and A.J. Caplan. 2003. "Willingness to Pay for Curbside Recycling with Detection and Mitigation of Hypothetical Bias." *American Journal of Agricultural Economics* 85(2):492–502.

Ariely, D., G. Loewenstein, and D. Prelec. 2003. "'Coherent Arbitrariness': Stable Demand Curves without Stable Preferences." *Quarterly Journal of Economics* 118(1):73–105.

Arrow, K., R. Solow, E. Leamer, P. Portney, R. Radner, and H. Schuman. 1993. "Report of the NOAA Panel on Contingent Valuation." *Federal Register* 58:4601–14.

Bjornstad, D., R. Cummings, and L. Osborne. 1997. "A Learning Design for Reducing Hypothetical Bias in the Contingent Valuation Method." *Environmental and Resource Economics* 10:207–21.

Blackburn, M., G.W. Harrison, and E.E. Rutström. 1994. "Statistical Bias Functions and Informative Hypothetical Surveys." *American Journal of Agricultural Economics* 76(5):1084–8.

Bohm, P. 1972. "Estimating the Demand for Public Goods: An Experiment." *European Economic Review* 3:111–30.

Brouwer, R., N. Powe, R.K. Turner, I.J. Bateman, and I.H. Langford. 1999. "Public Attitudes Toward Contingent Valuation and Public Consultation." *Environmental Values* 8:325–47.

Bulte, E., S. Gerking, J.A. List, and A. de Zeeuw. 2005. "The Effect of Varying the Causes of Environmental Problems on Stated WTP Values: Evidence from a Field Study." *Journal of Environmental Economics and Management* 49(2):330–42.

Carson, R., T. Groves, and M. Machina. 2000. "Incentive and Informational Properties of

Preference Questions." Working Paper, Department of Economics, University of California, San Diego.

Carson, R., T. Groves, J.A. List, and M. Machina. 2002. "Probabilistic Influence and Supplemental Benefits: A Field Test of the Two Key Assumptions Underlying Stated Preferences." Working Paper, Department of Economics, University of California, San Diego.

Champ, P.A., R.C. Bishop, T.C. Brown, and D.W. McCollum. 1997. "Using Donation Mechanisms to Value Nonuse Benefits from Public Goods." *Journal of Environmental Economics and Management* 33(2):151–62.

Cummings, R.G., G.W. Harrison, and L.O. Taylor. 1995. "Can the Bias of Contingent Valuation Surveys be Reduced? Evidence from the Laboratory." Working Paper, Department of Economics, Georgia State University.

Cummings, R.G., and L.O. Taylor. 1998. "Does Realism Matter in Contingent Valuation Surveys?" *Land Economics* 74(2):203–15.

——. 1999. "Unbiased Value Estimates for Environmental Goods: Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89(3):649–65.

Diamond, P.A., and J.A. Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" *Journal of Economic Perspectives* 8(4):45–64.

Fox, J., J. Shogren, D. Hayes, and J. Kleibenstein. 2003. "CVM-X: Calibrating Contingent Values with Experimental Auction Markets." *Experiments in Environmental Economics* 1:445–55.

Haab, T.C., and K.E. McConnell. 2002. *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*. Northampton, MA: Edward Elgar.

Harrison, G.W., and J.A. List. 2004. "Field Experiments." *Journal of Economic Literature.* 42(2):1009–1055.

Harrison, G.W., and E.E. Rutström. 2005. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." In *Handbook of Experimental Economics Results*.C. Plott and V.L. Smith eds., New York: Elsevier Science.

List, J.A. 2001. "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards." *American Economic Review* 91(5):1498–507.

List, J.A., and J. Shogren. 2002. "Calibration of Willingness-to-Accept." *Journal of Environmental Economics and Management* 43(2):219–33.

Loomis, J., T. Brown, B. Lucero, and G. Peterson. 1996. "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):450–61.

Loomis, J., A. Gonzalez-Caban, and R. Gregory. 1994. "Substitutes and Budget Constraints in Contingent Valuation." *Land Economics* 70(4):499–506.

Lusk, J.L. 2003. "Willingness to Pay for Golden Rice." *American Journal of Agricultural Economics* 85(4):840–56.

Poe, G., J. Clark, and W. Schulze. 1997. "Can Hypothetical Questions Predict Actual Participation in Public Programs? A Field Validity Test Using a Provision Point Mechanism." Working paper, Department of Agricultural and Resource Economics, Cornell University.

Powe, N.A., G.D. Garrod, and P.L. McMahon. 2005. "Mixing Methods within Stated Preference Environmental Valuation: Choice Experiments and Post-questionnaire Qualitative Analysis." *Ecological Economics* 52:513–26.

Smith, V.K. and C. Mansfield. 1998. "Buying Time: Real and Hypothetical Offers." *Journal of Environmental Economics and Management* 36(3):209–24.

Taylor, L.O. 1998. "Incentive Compatible Referenda and the Valuation of Environmental Goods." *Agricultural and Resource Economics Review* 27:132–9.