

SUGGESTED TITLE: Consequentiality and Contingent Values: An Emerging Paradigm

Gregory L. Poe
Department of Applied Economics and Management
Cornell University
422 Warren Hall
Ithaca, NY 14850-7801

and

Christian A. Vossler
Department of Economics
University of Tennessee
527C Stokely Management Center
Knoxville, TN 37996-0550

SUGGESTED TITLE: Consequentiality and Contingent Values: An Emerging Paradigm

In recent years a new paradigm has emerged with respect to the concept of “Hypothetical Bias” in contingent valuation. Following Bohm’s (1972) seminal public goods experiments, empirical criterion validity tests of contingent valuation have sought to compare hypothetical (“stated”) survey responses against the criterion of actual (“revealed”) economic commitments to public goods: “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” (Harrison and Ruström, 2008, p. 752). Whereas occasional research has found that hypothetical, stated values can be lower than actual commitments, reviews of hypothetical versus actual public goods contributions (e.g., Murphy and Stevens, 2004; Harrison and Ruström) and meta-analyses of these data suggest “that people tend to overstate their actual willingness to pay in hypothetical situations” (List and Gallet, 2001, p. 241; see also Little and Berrens, 2004 and Murphy *et al.*, 2005). These conclusions translate into what appears to be the conventional wisdom regarding contingent valuation (CV): “A fundamental concern of any CV study is hypothetical bias. Respondents have a well-established tendency to state willingness to pay values that are significantly greater than those revealed in real-market interactions” (Aadland *et al.*, 2007).

In a series of conference presentations, working papers, and a journal article, Richard Carson and Theodore Groves and co-authors (e.g. Carson and Groves *et al.*, 1997, 1999, and 2007) present an alternative paradigm for conceptualizing the relationship between stated preference survey responses and real economic commitments to private and public goods. To paraphrase Kuhn (1970, p. 81), in his influential treatise, *The Structure of Scientific Revolutions*, these authors have looked at the same stated preference data, but placed them in a new set of

relations with one another by giving them a different framework. Specifically, they argue that the bifurcation of data into purely hypothetical responses and real actions is misplaced and uninformative from an economic perspective. Whereas psychologists have developed theories of hypothetical responses (e.g., Kahneman *et al.*, 1982), “economic theory has nothing to say about...purely hypothetical questions” (Carson *et al.*, 2004). Building on the mechanism design literature, economic theory does, however, offer a predictive, theoretical framework for interpreting responses that have the potential to influence agency action.

In this chapter, we summarize the theoretical arguments of Carson and Groves, *et al.*, and assemble early empirical evidence that comports with this theoretical framework. In doing so, we argue that redefining criterion validity in terms of consequentiality offers the potential for a fundamental paradigm shift in the Kuhnian sense. That this shift has yet to be fully incorporated into the contingent valuation literature reflects the nascent state of this paradigmatic challenge as well as the continued inertia of the dominant hypothetical bias paradigm. Further, empirical support for Carson and Groves, *et al.*’s consequentiality arguments have emerged in a somewhat piecemeal manner, spread across a diverse set of journal articles and unpublished manuscripts,

Consequentiality: Conceptual Framework

In this section we liberally draw from two sets of papers that lay the conceptual foundations of the consequentiality framework to criterion validity in contingent valuations (Carson and Groves *et al.* 1997, 1999, 2007; and Carson *et al.*, 2004). The critical component of these is captured in the following comparative definitions.

“Consequential survey questions: If a survey’s results are seen by the agent as potentially influencing an agency’s actions and the agent cares about the outcomes of those actions, the agent should treat the survey questions as an opportunity to influence those actions. In such a case, standard economic theory

applies and the response to the question should be interpretable using mechanism design theory concerning incentive structures.

Inconsequential survey questions: If a survey's results are not seen as having any influence on an agency's actions or the agent is indifferent to all possible outcomes of the agency's actions, then all possible responses by the agent will be perceived as having the same influence on the agent's welfare. In such a case, economic theory makes no predictions." (Carson and Groves, p. 183)

The authors argue that in responding to a consequential survey question "a rational economic agent will take the incentive structure of a consequential survey into account in conjunction with information provided in the survey and beliefs about how that information is likely to be used" (Carson and Groves, p. 204). Cover letters accompanying surveys typically stress that everyone's response matters, in part to maximize response rates. Survey instruments generally suggest that responses will inform the policy process, and increase realism by, for example, providing reminders of substitute goods and budget constraints. This does not imply that respondents to a consequential survey, however, will necessarily reveal their true preferences. Indeed they may respond strategically if the question format is not incentive compatible, e.g. if it is an open-ended format. Also, even with an incentive-compatible elicitation, if respondent beliefs about the proposed outcome or its cost upon implementation differ from what is stated in the survey, the analyst will not be able to recover the respondent's true preferences without knowledge of these beliefs. In this case, at least theoretically, the elicitation yields a truthful response but to a different proposal.

Foundations: Incentive Compatibility and Demand Revelation of Binary Referendum Questions

The incentive structure used in a survey question is therefore critical to the theoretical formulation and empirical evaluation of consequentiality. When moving from theoretical conceptualizations to empirical explorations, it is useful to distinguish between the theoretical

notion of incentive compatibility and the empirical notion of demand revelation. In many presentations these concepts have been used interchangeably (e.g. Cummings *et al.*, 1995; Taylor *et al.*, 2001) with the lack of clarity being a source of confusion. Here we are explicit in our terms. An incentive-compatible mechanism is a theoretical concept, meaning that a respondent has an incentive to truthfully reveal his preferences. Demand revelation is an empirical concept, providing a measure of how well decisions correspond with true, underlying values. If other conditions are not satisfied, it is possible that a theoretically incentive-compatible mechanism is not demand revealing, and vice versa.

Building upon mechanism design theory, most specifically what is referred to as the Gibbard-Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975), Carson *et al.* (2004) argue that binary referenda can be incentive compatible under the following conditions.

“Proposition 1: A binding (binary) referendum vote with a plurality aggregation rule is incentive compatible in the sense that truthful preference revelation is the dominant strategy when the following additional conditions hold: (a) the vote is coercive in that all members of the population will be forced to follow the conditions of the referendum if the requisite plurality favors its passage and (b) the vote on the referendum does not influence any other offer [that] might be made available to the relevant population.”

Four recent studies (Taylor *et al.*; Vossler and McKee, 2006; Burton *et al.* 2007; Collins and Vossler, 2009)¹ have used induced value laboratory experiments to explore the demand revelation characteristics of referenda that correspond with Carson *et al.*'s (2004) Proposition 1. Following Smith (1976), assuming that individuals abide by the postulate of non-satiation and are otherwise rational, experimental preferences can be achieved in a controlled economic laboratory setting by using a reward structure to induce prescribed monetary value on actions.

¹ In the analysis that follows, we use data from Vossler and McKee's dichotomous choice “real, certainty” treatment, Burton *et al.*'s “consequential” treatment for the three experiments they report on, and the dichotomous choice “plurality” treatment data from Collins and Vossler.

The basic idea is that given the opportunity to choose between two alternatives, “identical except that the first yields more of the reward medium (usually currency) than the second, the first will always be chosen (preferred) over the second.” (Smith (1976), p. 275). Building on this idea, the common framework for exploring demand revelation in these experiments is that individuals are offered the opportunity to vote for a public good that will be provided to all individuals in the group if a specified voting threshold is surpassed. The public good has value in the sense that it provides personal monetary rewards to each individual i in the group. If the provision rule is met individuals have to pay a specified cost, C_i , for its provision and each individual receive his or her induced value, V_i .² If the voting threshold is not achieved, no one receives any payment and no one incurs any cost: the public good is not provided. As an example, Taylor *et al.* replicate the Carson *et al.* (2004) Proposition 1 conditions in a secret ballot, majority rule referendum which they describe as follows:

“If it passed, each subject in the room would pay \$5 (regardless of whether they voted yes or no) and in return, they received ‘the good’, which was simply the amount of money that would be paid to them at the end of the experiment. If it did not pass, no one paid \$5 and no one received the good, regardless of how they voted... They were instructed to vote ‘yes’ if they would like the referendum to pass, and ‘no’ if they did not want the referendum to pass.” (p. 63)

Participants were informed of their personal induced value and told that not everyone had the same value. They were not however informed of the range of values or how the values were distributed across other participants.

In the Taylor *et al.* experiment, the personal induced value of the good varied across participants, while all individuals paid the same cost (i.e., $C_i=5$ for all i). Experiments 1 and

² Collins and Vossler’s experiment is couched in a choice experiment elicitation framework. To make this data compatible with the other studies, the cost is taken to be the net benefits associated with the status quo choice or “no” vote whereas the value is the net benefits associated with the alternative choice or “yes” vote.

2 of the Burton *et al.* study follow a similar design.³ The other induced-value experiments varied both costs and benefits: Vossler and McKee used induced values over the range from \$1.50 to \$9.50, in \$1 increments, and costs of \$1, \$3, \$5, \$7 and \$9; in Burton et al.'s induced values in Experiment 3 were \$4 and \$8 while costs were \$3, \$5, \$7 or \$9; for the dichotomous choice component of the Collins and Vossler study, induced values range from -\$1 to \$10.50, in \$0.50 increments, with associated costs of \$2, \$3, \$4 and \$5. While the underlying experimental designs vary across the four studies, a common feature is that each can be represented in terms of the distribution of "No" responses as a function of induced personal value minus personal costs (i.e., $V_i - C_i$).

The $V_i - C_i$ cumulative distribution functions (CDFs) for the five groups of experiments are presented in Figure 1.⁴ If the referendum mechanism is perfectly demand revealing, we expect 100 percent 'yes' votes for negative difference values and zero percent 'yes' votes for positive difference values. In other words there would be single step from 100 to 0 percent at $V_i - C_i = 0$. It is evident from the graphs that perfect demand revelation is not observed in any of the four studies. The four studies combined yield 556 of 603, or 92.2%, of votes consistent with induced preferences. As a test of demand revelation, we use one-sample Kolmogorov-Smirnov (K-S) tests to test the null hypothesis that the empirical

³ Burton *et al.* conduct three experiments "using two very different participant groups (cadets at a US military academy and university students in Northern Ireland)" (pps. 518-519), and report finding different behaviors across the two groups. On this basis we group Experiments 1 and 3, conducted using cadets from the US military academy, and treat Experiment 2, with the university students from Northern Ireland, separately. The basic findings reported in the text for this grouping hold for the alternate in which each experiment is treated separately.

⁴ To obtain a valid CDF we imposed a monotonicity constraint following the approach described by Haab and McConnell (1997).

and theoretical difference value CDFs are equal.⁵ This null hypothesis is rejected at the 5% level for three of the five data sets, Taylor *et al.*, Burton *et al.* (Experiment 2) and Collins and Vossler. However, for two of the three studies where K-S tests reject the hypothesis of perfect demand revelation the “mis-votes” or digressions from induced-preferences at the source of the rejection occur at rather small difference values: for Taylor *et al.* this is between -\$0.10 and \$0.50; for Vossler and McKee this is between \$0 and \$0.50. As Taylor *et al.* argue, it may be that “rewards and penalties of less than \$1 may not have been salient” (p. 65). Similarly, along the line of thought underlying random utility modeling, the assumption that utility is driven by induced values may only capture the true underlying utility comparison with some degree of unobservable error. For example, social preferences may play a role. Models relying only on induced values do not capture possible altruism or alternative forms of other-regarding preferences participants may have. We do not, however, believe that these arguments carry over to the level of “mis-votes” observed in the Burton *et al.* Experiment 2 for more salient difference values of -\$2 and \$2. For the moment at least, this Experiment appears to be an unexplained outlier, in which behavior for this group of students seems differs considerably and significantly from the economic predictions as well as from the results of the body of experiments in this area.

We further note that although there are some systematic differences between empirical and theoretical difference-value distributions, this does not necessarily translate into biased willingness to pay (WTP) distributions. That is, it may be the case that at a

⁵ The *D*-statistic corresponding with the K-S goodness of fit test is the absolute value of the maximum difference between the theoretical and empirical distributions. For a large sample, the critical value for this test associated with a 5 percent significance level is approximately $1.36\sqrt{n}$, where *n* is the sample size of the empirical distribution.

particular cost amount that ‘yes’ and ‘no’ ”mis-votes” cancel each other out to a large degree. We turn to an analysis of “aggregate” demand revelation below.

In a similar fashion to our analysis of difference values, we compare empirical and theoretical WTP CDFs using K-S tests for the Vossler and McKee and the Collins and Vossler studies^{6,7}. To be clear, these CDFs are constructed from the proportion of “no” responses to each cost amount, rather than the proportion of “no” responses to each difference value. We fail to reject the null hypothesis of equality between empirical and theoretical WTP distributions in both studies.

Taken together, we interpret the results from the above tests as indicative that decisions are consistent with aggregate demand revelation, but there are some deviations from induced preferences at particular dollar values. Most of the deviations from induced preferences occur for small differences between induced value and cost. As mentioned above, uncontrolled social preferences may explain at least some of these decisions.

Demand Revelation in Binary Referendum with Uncertainty in Values

⁶ An alternative to the approach used in the text would be to estimate mean and median values for the difference distributions presented in Figures 1 and 2, and to compare these estimates to the theoretical prediction that the mean and median values of the $V_i - C_i$ distribution are equal to zero. Rejection of this test would suggest that errors are systematic in one direction relative to $V_i - C_i = 0$. We do not adopt this approach here because the validity of the test would rest on the assumption that the V_i and C_i are not correlated. Based on our personal knowledge of the Vossler and McKee study, this is not the case. In the Vossler and McKee study the V_i and C_i values were deliberately correlated. Personal communication with Mike McKee indicates that the V_i to C_i relationship in the Taylor *et al.* study was specifically structured so that it was unlikely that the referenda would pass, thus facilitating more rapid payouts in classroom experiments.

⁷ The Taylor *et al.* and Experiments 1 and 2 of Burton *et al.* have a single costs. The values and costs of the third experiment in Burton *et al.* had value to cost relationships that were, for some costs only one-sided, not allowing for a two sided error distribution. For these reasons we judged that the data from these studies was not amenable to estimating willingness-to-pay distributions.

To this point, our discussion has only reported results from demand revelation studies in which the induced values are certain. It has, however, long been recognized (e.g. Opaluch and Segerson, 1989; Dubourg et al., 1994; Ready *et al.*, 1995) that rather than having a single point estimate of the value for the environmental or public good, contingent valuation respondents may instead have a distribution or range of possible WTP values. To explore the effect of uncertainty on demand revelation of the binary referendum mechanism, Vossler and McKee built upon their previously discussed experimental design by inducing uncertainty in values as follows:

“For value certainty treatments, induced values across group members are uniformly distributed over the range of \$1.50 to \$9.50, in \$1 increments. For uncertain value treatments, participants are given a \$2 range of possible values. These ranges are constructed by adding/subtracting \$1 from the set of certain values. This range is wide relative to the value distribution. Participants are instructed that each value in the range (in 25-cent increments) has an equal chance of being selected. After all the [voting] decisions are made, the exact value for each participant is determined through a die roll” (p. 142).

Under the assumption that individuals are expected utility maximizers that base their voting decisions on the expected difference value, $E[V_i - C_i] = E[V_i] - C_i$, Figure 2 plots the distribution of “No” responses across expected induced difference values using the methods described above. Using a K-S test, we reject the null hypothesis of equal empirical and theoretical difference-value functions at the 5% level. Note, however, that this rejection is driven by votes made under expected differences of $-\$0.50$ and $\$0.50$, both of which lie within the range of induced uncertainty. Arguably, if respondents invoke decision heuristics other than expected value maximization, it is not valid to assume that these represent errors. The null hypothesis of equality between the theoretical and empirical distributions cannot be rejected when differences from $-\$0.50$ to $\$0.50$ are excluded. Similar to the above results with certain values, we do not find any bias between the theoretical and empirical WTP functions. While the evidence regarding uncertain induced values is limited to one study,

these results are consistent with the notion that the demand revelation characteristics of the incentive-compatible consequential binary referendum carry over to cases for which induced values are uncertain.

Framed Field Experiments: Homegrown Values and Consequentiality

While induced-value laboratory experiments provide critical information about the demand revelation characteristics of consequential, incentive-compatible value elicitation mechanisms for public goods, contingent valuation is inherently a field method to elicit “homegrown values” that an individual might have for a nonmarket environmental or public good. “Homegrown value”,

“refers to a subject’s value that is independent of the value which an experimenter might ‘induce’ (see Vernon L. Smith, 1976). The idea is that homegrown values are those that the subject brings *to* an experiment” (Cummings *et al.*, p. 260)

Researchers have used framed field experiments, which differ from a conventional laboratory experiment in a number of ways, to further explore consequentiality. As defined by Harrison and List (2004) laboratory experiments conventionally use a standard participant pool of students, frame the decision abstractly and impose a set of rules. A framed field experiment instead uses a non-standard participant pool with a “field context in either the commodity, task, or information set that the subjects can use” (Harrison and List (2004), p. 1014).

Landry and List (2007) and Carson *et al.* (2004) undertake a series of framed field experiments using participants from a well functioning marketplace – the floor of a sports card show in Tucson, Arizona. Participants were recruited as they entered the show for a public goods experiment run in a separate room in the same building. The public good is the

provision of n identical pieces of sports memorabilia if the majority of n participant votes to fund “Mr. Twister.” An excerpt from Landry and List describes the good.

“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.” (p. 423)

The \$X values were \$5 and \$10 in Landry and List. In the Carson *et al.* (2004) study, “Mr. Twister” distributed Kansas City Royal game ticket stubs dated June 14 1996, which were issued for admission to the baseball game in which Cal Ripken Jr. broke the world record for consecutive games played. The cost to each individual of funding “Mr. Twister” was \$10.

We report on three treatments here. First, in what we label as the “baseline” treatment the ticket stubs were provided and everyone paid the indicated costs if the majority of the participants voted to fund “Mr. Twister”. If 50% or less of the participants voted to fund “Mr. Twister”, no one paid the fee and no one received a ticket stub. The “probabilistic referenda” was the same as the baseline treatment, with the following exception. If the majority voted to fund “Mr. Twister”, then a second step, a coin flip, would be used to determine if the funding decisions would be binding. The funding decision was binding if the coin flip turned up heads, a 50% probability. To impose other probabilities, a 10-sided die was used in Carson *et al.*(2004). If, for example, a 20% chance was being used, and the die turned up one or two, the ticket stubs would be provided and all participants would have to pay the specified amount. If the die turned up a value between three and 10, “Mr Twister” would not be funded. In the “hypothetical” treatment, “passive language was used so that subjects understood that their vote would not

induce true economic consequences – i.e. no money would change hands” (Landry and List, p. 423).

Referring back to the consequential/inconsequential definitions, and Proposition 1 above, the baseline treatment and the probabilistic referenda satisfy the conditions for an incentive-compatible elicitation mechanism. However, in contrast with the induced-value laboratory experiments it is not possible to test for demand revelation by comparing induced-value and cost distributions. Instead, with homegrown-value criterion validity studies, a common approach is to use results from an incentive-compatible elicitation as a benchmark from which to compare treatments intended to more closely capture the contingent valuation setting. As a result, the relevant null hypothesis here is simply that each of the probabilistic treatments results in vote proportions that are equal to those in the baseline, binding referenda. In contrast, the hypothetical treatment is inconsequential and hence, we are unable to formulate economic-theoretic expectations of voting patterns vis-à-vis the baseline and probabilistic referenda.

Selected relevant results from these two studies are reported in Table 1. Examination of the Table suggests that there is little difference in the voting behavior between the baseline treatment and the probabilistic referenda. However, the proportion of ‘yes’ votes in the hypothetical treatment is considerably higher than the consequential treatments. Statistical tests of the hypotheses confirm that the distribution of voting decisions is equal amongst the consequential treatments, i.e. the null hypothesis of equality cannot be rejected. However, the equality of the voting behavior between the consequential and inconsequential (i.e., hypothetical) treatments can be rejected. Interestingly, the difference in distributions appears to be “knife edged”: even low-probability referenda provide similar values as the baseline treatments.

The lesson from these data is best summarized by Landry and List:

“...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.” (p. 427)

Advisory Referenda

The framed field test experiments support the hypothesis that as long as decisions can probabilistically influence an outcome, individuals have incentives to respond as if the referenda were binding. This notion of probabilistic referenda, however, deviates from the situation presented in most contingent valuation studies. Rather, “cover letters for SP studies often state that the survey results will be shared with state or local officials....Survey instruments generally provide additional signals that respondents should take seriously the valuation exercise” (Vossler and Evans, *forth.*). In this manner, the aspect of potential consequentiality is stressed, but not probabilistically. Rather, the survey responses are presented as being advisory to decision makers. That such efforts by survey researchers are effective is evidenced in a recent contingent valuation study of willingness to vote in favor of a referendum to improve water quality at an Iowa lake by Herriges *et al.* (*forth.*). This survey included a Likert-scale question to measure the respondents’ belief about the likelihood that survey results would affect policies related to water quality in Iowa lakes. A “one” response indicated “no effect at all” and a “five” response denoted “definite” effects. Less than 7 percent of those returning a survey reported a value of one, suggested that only a small proportion of respondents regarded the survey as being inconsequential.

Carson *et al.* (2004) address this advisory nature of contingent valuation research in a separate proposition:

“Proposition 2: changing from a binding referendum to an advisory referendum doesn’t alter the incentive structure as long as the decision maker is more likely to undertake the referendum proposed outcome if the specified plurality favors it. This proposition follows from noting that it is the nature of the influence on the decision (the agent is potentially pivotal at one point in the decision space, the requisite plurality, with only a binary weight on the in the aggregation rule) not the binding nature of the referendum that matters.”

In general, incentive compatibility in the advisory referendum will be based on the respondents’ perceived influence on the outcome.

In a homegrown value laboratory experiment, Vossler and Evans used a split sample design to compare responses to binding, advisory and hypothetical referenda. Student participants were asked to vote in a referendum on whether everyone in the group (session) would fund the provision and administration of one on-campus, classroom recycling container at a particular cost. Consistent with Proposition 1(a), the funding mechanism was coercive. In addition, there was no clear venue for which students themselves can purchase recycling containers (and have them maintained by the university), such that the binding and advisory referenda are likely to also satisfy Proposition 1(b).

The binding referendum, as in the prior studies, serves as a baseline for comparison. It involved a referendum with majority-vote implementation rule, which has already been noted to be incentive compatible and demonstrated to be demand revealing in induced-value experiments. In the advisory referendum, which Vossler and Evens refer to as an “implicit advisory referendum”, efforts were made to make the instructions as close as possible to a field contingent valuation survey in which there is no direct signal on exactly how responses will be used in a policy decision. In this treatment, participants were given the following information.

“Passage of the referendum will not solely be determined by how you and the other participants vote. In particular, we, the experiment coordinators, will use your votes as *advice* on whether or not to pass the referendum. While you will not be told how we came to a decision, know that the likelihood the referendum is passed increases with the number of YES votes cast.” (Vossler and Evans, *forth.*)

In actuality, unknown to the participants, the decision rule used was identical to the baseline, i.e. a majority-rule vote with no experiment coordinator votes. The hypothetical treatment framed a referendum similar to the baseline, but with slightly different language to make clear that the vote was inconsequential.

The results of the three treatments are provided in Figure 3, which depict the probability of a ‘yes’ response for costs of \$1,\$3,\$6 and \$8. As depicted, the distribution of votes in the hypothetical treatment lies to the right of the other two treatments. Using two-sample K-S tests, the null hypothesis of equal WTP distributions is marginally rejected between the hypothetical and baseline treatment ($D=0.231, p < 0.10$) and between the hypothetical advisory vote ($D=0.231, p < 0.10$). The null hypothesis of equality between the baseline and the advisory distributions cannot be rejected ($D=0.104, p > 0.10$).⁸ These results are only suggestive however, as with homegrown values there is no guarantee that the underlying true WTP distribution is the same across treatments. Using parametric models that account for whether the student has a class in the building designated as the location of the proposed recycling container, and socio-economic variables, the authors find that elicited WTP in the hypothetical referendum is statistically different, and is roughly 100% higher, than in the baseline. Yet there is no statistical difference in elicited WTP between baseline and advisory referenda.

Based on their statistical analyses, Vossler and Evans conclude that the results of their experiments “designed to capture key characteristics of a SP survey for a proposed

⁸ In their paper, Vossler and Evans test whether the overall proportion of ‘yes’ responses are equal across treatments using Fisher exact tests, which yields stronger evidence of hypothetical bias.

environmental program, provide support for the theoretical predictions regarding voter behavior in advisory referenda.” This conclusion is further supported in the Herriges *et al.* comparison of the estimated WTP distributions of respondents who indicated that the Iowa lake survey was inconsequential and respondents who indicated otherwise.

“we find support for the equality of WTP distributions among those believing the survey is at least minimally consequential, while those believing the survey will have no effect on policy have statistically different distributions associated with WTP” (Herriges *et al.*, *forth.*).

These results are consistent with the “Mr Twister” framed field experiments.

A Contingent Valuation Criterion Validity Test

The last empirical piece of evidence in support of the Carson and Grove’s *et al.*’s theory of consequentiality is a contingent valuation criterion validity test conducted by Johnston (2006).

This study compares:

“... genuine discrete choice CV responses to aggregated votes in a subsequent, binding public referendum. The assessment is designed to be unambiguous and simple. hypothetical and actual choice contexts are parallel and consequential, and address the provision of an identical quasi-public good (i.e., the provision of public water to the Village of North Scituate, Rhode Island). Respondents are drawn from the same well-defined population. No re-coding or transformation of survey responses is required, no cheap-talk or certainty adjustments are applied, and a “one vote per survey” format eliminates the need to adjust for correlation or sequence effects among responses.” (p. 470)

The contingent valuation study was conducted to assist the Village committee assess the public support for a public water provision project. The intent was to use survey methods to determine if there should be an officially sanctioned referendum on the project, as required by the State of Rhode Island. Sanctioning, promoting, scheduling and implementation a referenda incurs significant costs. Johnston continues:

“Although the survey instrument noted the possibility of a public vote as a possible subsequent step in the process of establishing the water supply project, this was the first indication that any official referendum might be forthcoming. As the survey was designed as a means to assess public preferences—before the official vote was approved or scheduled—it provides a nearly ideal context in which to assess the validity of hypothetical survey responses in a genuine CV context.” (p. 472).

Based on the contingent valuation study, the village decided to pursue a real referendum, wherein the quarterly cost per household was estimated to be \$250.

The results from this comparison of an advisory contingent valuation survey and a real vote are consistent with the previously discussed laboratory and field experiments. As depicted in Figure 4, the contingent valuation response distribution across a number of possible costs closely predicts the actual proportion of “Yes” responses to \$250 in the real referendum. The null hypothesis of equality between the contingent valuation (48.4%) and the actual (45.7%) ‘yes’ vote percentage at \$250 can not be rejected ($p=0.69$). As in Vossler and Evans, these results are further supported by parametric estimations of the contingent valuation WTP distribution.

Discussion

With few exceptions, conclusions from past criterion validity studies cast contingent valuation in a much different light than studies that focus on other aspects of validity. First, there is strong evidence of construct validity: consistent with consumer demand theory, elicited contingent values vary with factors such as income and scope, and elicited willingness to accept exceeds WTP (see Carson, 1997; Carson *et al.*, 2001). Second, there is strong evidence of convergent validity: estimates of value from contingent valuation studies approximate those from revealed preference studies (see Carson *et al.*, 1996). This suggests that, at least for eliciting use values, contingent valuation may be appropriate. Third, there are persistent elicitation effects

demonstrating that, consistent with expectations from mechanism design theory, the mechanism used to elicit contingent values matters. In particular open-ended elicitation questions lead to lower estimates of value than a dichotomous choice (e.g., Cameron *et al.*, 2002). These elicitation effects provide further evidence of construct validity.

So then, why is there a divergent conclusion from the majority of criterion validity studies? We hypothesize that reliance on using a purely inconsequential decision setting as the analog to a stated preference survey is at least a partial explanation. Based on the results from the studies highlighted in this chapter, if survey respondents perceive their decisions to be consequential, then this motivates responses that are quite different from those in inconsequential settings. As such, placing the results of criterion validity studies in their proper context of consequentiality can help reduce misinterpretations. On a related note, criterion validity studies that compare *consequential* surveys with actual behavior need also cast results in their proper context by drawing on the foundations put forth by Carson and Groves, *et al.* In many cases, for example, these studies compare voluntary contributions with consequential surveys cast in a similar context (e.g., Champ *et al.*, 1997). In such a framing Proposition 1(b) is likely violated and hence, as argued by Carson and Groves, those in the contingent setting should overstate their value if they believe doing so will increase the chance of an actual fundraising campaign being implemented. On the other hand, as shown in Poe *et al.* (2002) the voluntary nature of the actual contributions in such comparisons will engender free riding, further muddying any efforts to compare actual contributions and purely inconsequential survey responses. Rather than providing a test of criterion validity, the observed deviation in response patterns, if they are consistent with

mechanism design theory, may be construed as empirical evidence of construct validity rather than as evidence that contingent valuation lacks criterion validity.

Our discussion is not meant to imply that we should reject the accumulated, sizable literature that explores hypothetical bias. As suggested by the Herriges *et al.* study, where there was a history of similar surveys affecting public policy, not all respondents indicated that they perceived the survey to be consequential. In other settings, such as one where respondents would question the ability of the authority to coercively collect payment, perceive that the agency would place zero weight on public opinion, or for what appears to be a purely academic study, there are likely to be a much higher proportion of respondents for which consequentiality does not hold.

Assuming there is a suitable way to separate respondents into consequential and inconsequential camps, researchers are still charged with the task of uncovering the demand of those in the latter group. Cheap talk scripts, alternative elicitation formats and/or calibration methods will continue to be essential in this regard. Justification for continued reliance on these approaches include minimizing possible sample selection bias that may otherwise arise, as well as the maximizing the number of useful surveys collected for a given budget.

We further note that there may be a connection between consequentiality and various aspects of survey design, including the use of cheap talk and language in cover letters. For instance, cheap talk scripts serve to emphasize the importance of obtaining accurate signals of value, and may hence increase the proportion in the consequential camp. On a related note, consequentiality may confound split-sample field survey comparisons if there is uncontrolled correlation between treatment and consequentiality. As such, this is yet one more reason why the

development and use of questions to elicit perceptions about consequentiality remain an important area of research.

From the scant number of methodological studies that are grounded in mechanism design theory, it is clear that there is an abundance of additional, fundamental research questions that remain. First, with the exception of Johnston, past research on consequentiality has been comprised of very controlled studies with attributes that differ from the field survey environment, making further field research a priority. Within the laboratory setting the experimental design may be modified to invoke closer correspondence with the field survey setting. Further, given that many researchers, usually out of concern for statistical efficiency or scenario plausibility, continue to use alternatives to the referendum elicitation format, sometimes with a voluntary contributions payment vehicle, exploration of these formats in controlled but consequential decision settings is warranted.

On a final note, there is a very pragmatic reason for redefining criterion validity in terms of consequentiality: we, as stated preference researchers, have been our own worst enemy in the debate over the criterion validity of contingent values. Stated preference researchers who cast surveys as purely inconsequential exercises are providing the impetus to other economists, academics and policy makers to dismiss results from our work. By grounding our work in mechanism design theory we, like many other economists, would have theoretical justification for what we do as well as a framework from which to undertake empirical tests. Pursuing this path will no doubt lead to theoretical refinements that will not only inform stated preference research, but more broadly the economics profession in areas such as decision making under uncertainty and voting. Further, shifting to the consequentiality paradigm should serve to increase the demand for stated preference research in the private and public sectors.

Figure 1. Proportion of “No” Respondents, Difference Value

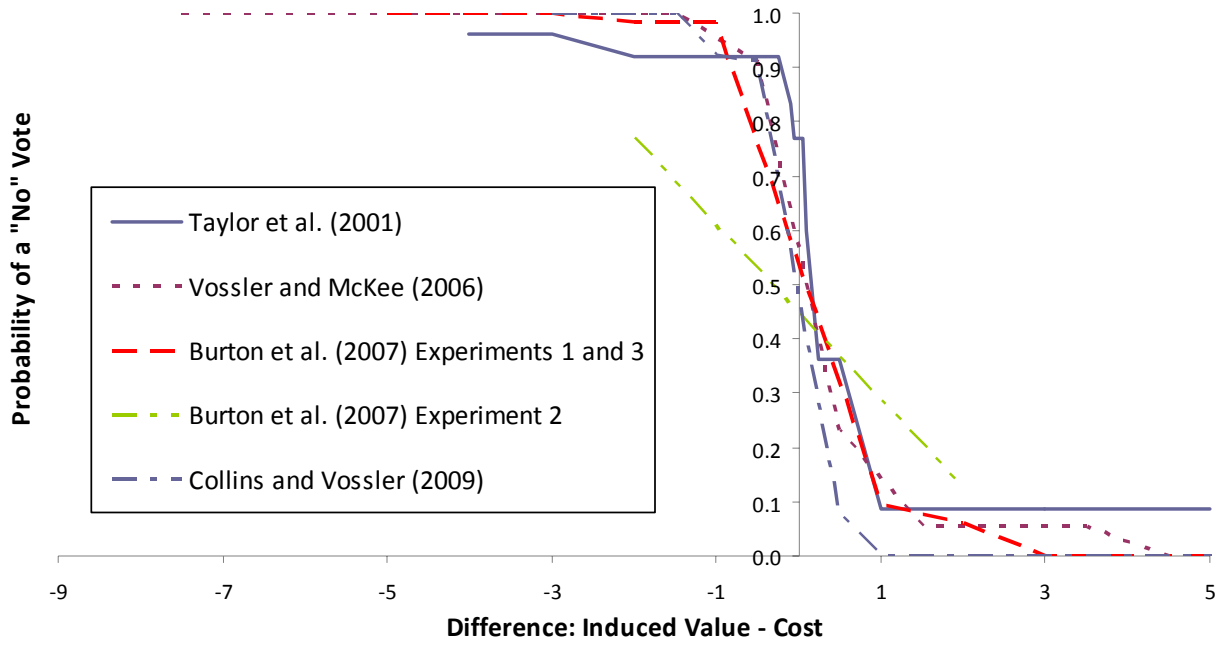


Figure 2: Proportion of “No” Respondents, Expected Difference Value, Vossler and McKee (2006)

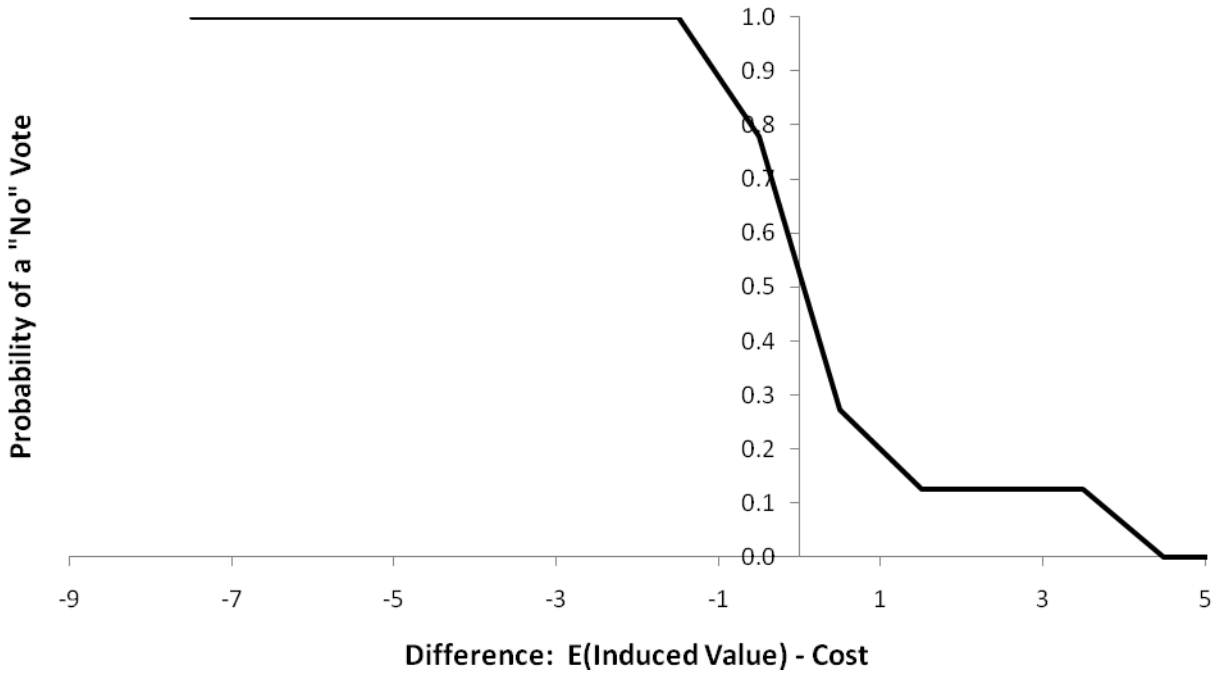


Figure 3: Real, Advisory and Hypothetical Voting Distributions, Vossler and Evans (forth.)

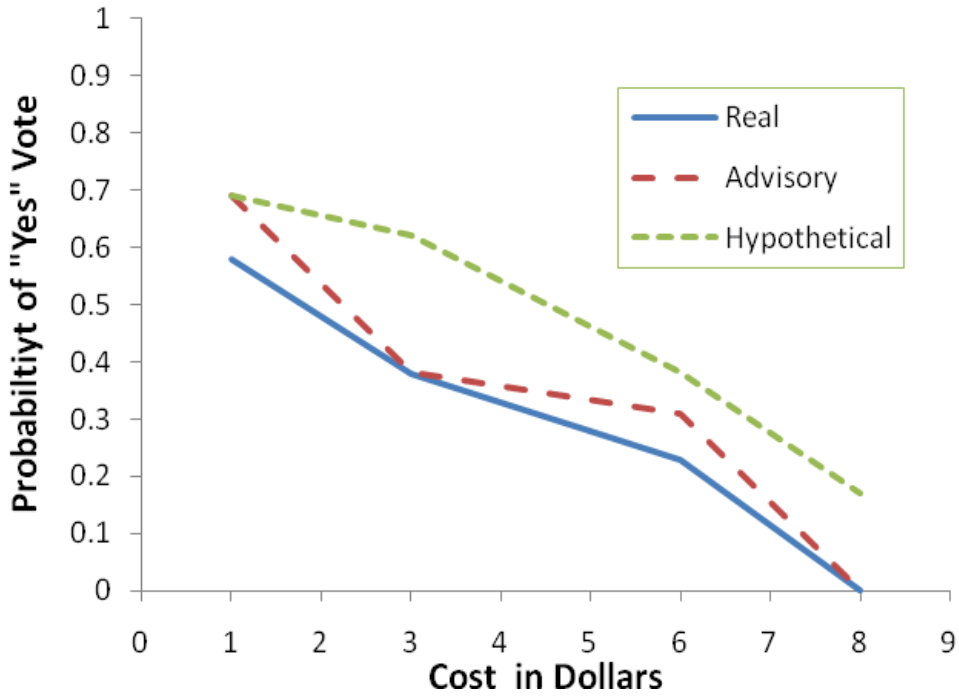


Figure 4: Contingent Valuation and Real Referendum Comparison, Johnson (2006)

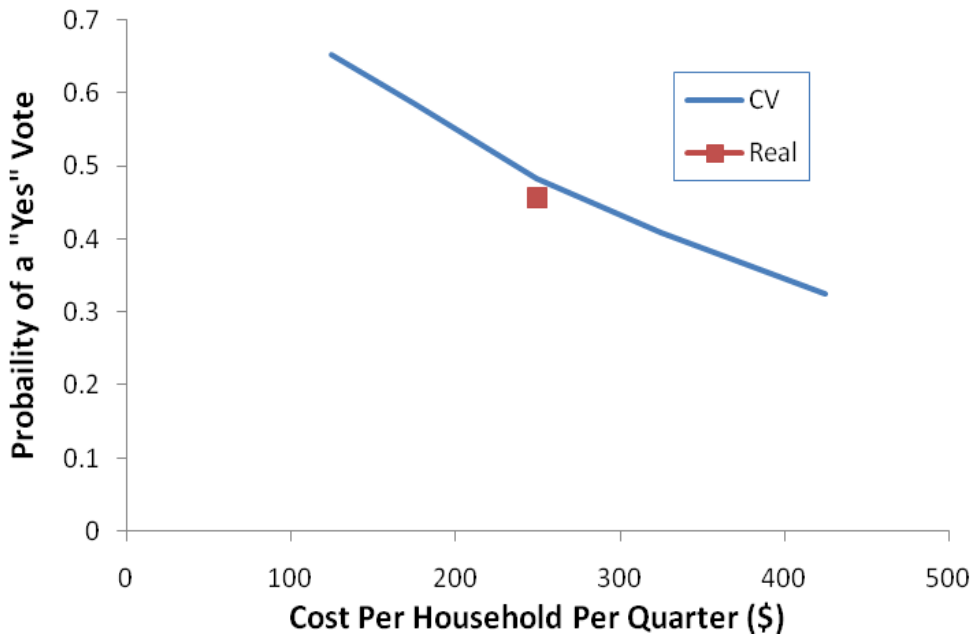


Table 1: Field Tests of Consequentiality: Percent “Yes” (number of observations)

Description	Consequential			Inconsequential	
	Real	Probabilistic % Chance of Being Binding		Hypothetical	
		0.80	0.50		0.20
Landry and List (2007), \$5	33 (64)	n.a.	32 (59)	n.a.	84 (64)
Landry and List (2007), \$10	19 (64)	n.a.	20 (59)	n.a.	75 (64)
Carson <i>et al.</i> (2004), \$10	46 (96)	41 (46)	48 (52)	44 (50)	60 (58)

References

- Aadland, D., B. Anatchkova, B. Grandjean, J.F. Shogren, B. Simon and P.A. Taylor (2007), "Valuing Access to Our Public Lands: A Unique Public Good Pricing Experiment", Selected Paper Presented at the American Agricultural Economics Association meetings, Portland OR, July.
- Bohm, P. (1972), "Estimating Demands for Public Goods: An Experiment", *European Economic Review*, **3**: 111-130.
- Burton, A.C., K.S. Carson, S.M. Chilton and W.G. Hutchinson (2007), "Resolving Questions about Bias in Real and Hypothetical Referenda", *Environmental and Resource Economics* **38**(4): 513-525.
- Cameron, T.A., G.L. Poe, R.G. Ethier and W.D. Schulze (2002), "Alternative Non-Market Value-Elicitation Methods: Are the Underlying Preferences the Same?", *Journal of Environmental Economics and Management* **44**: 391-425.
- Carson, R.T. (1997), "Contingent Valuation Surveys and Tests of Insensitivity to Scope", In R.J. Kopp, W. Pommerhene and N. Schwartz, eds., *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods*. Boston: Kluwer, pp. 127-163.
- Carson, R.T., N.E. Flores, K.M. Martin and J.L. Wright (1996), "Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods", *Land Economics* **72**(1): 80-99.
- Carson, R.T., N.E. Flores and N.F. Meade (2001), "Contingent Valuation: Controversies and Evidence", *Environmental and Resource Economics* **19**(2): 173-210.
- Carson R.T., T. Groves and M.J. Machina (1997), "Stated preference questions: context and optimal response." Paper presented at the National science foundation preference elicitation symposium, University of California, Berkeley.
- Carson R.T., T. Groves and M.J. Machina (1999), "Incentive and Informational Properties of Preferences Questions", Plenary Address, European Association of Environmental and Resource Economists, Oslo, Norway.
- Carson, R.T and T. Groves (2007), "Incentive and Informational Properties of Preference Questions", *Environmental and Resource Economics* **37**(1): 181-210.
- Carson, R., T. Groves, J. List and M. Machina (2004), "Probabilistic Influence and Supplemental Benefits: A Field Test of Two Key Assumptions Underlying Stated Preferences", Unpublished Draft Manuscript.
- Champ, P.A., R.C. Bishop, T.C. Brown and D.W. McCollum (1997), "Using Donation Mechanisms to Value Non-Use Benefits from Public Goods", *Journal of Environmental Economics and Management* **33**(2): 151-163.
- Collins, J.P. and C.A. Vossler (2009), "Incentive Compatibility Tests of Choice Experiment Value Elicitation Questions", *Journal of Environmental Economics and Management* **58**(2): 226-235.
- Cummings, R.G., G.W. Harrison and E.E. Ruström (1995), "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive Compatible?", *American Economic Review* **85**(1): 260-266.
- DuBourg, W.R., M.W. Jones-Lee and G. Loomes (1994), "Imprecise Preferences and the WTP-WTA Disparity", *Journal of Risk and Uncertainty* **9**: 115-133.

- Gibbard, A. (1973), "Manipulation of Voting Schemes: A General Result", *Econometrica* **41**(3): 587-602.
- Haab, T.C. and K.E. McConell (1997), "Referendum Models and Negative Willingness to Pay: Alternative Solutions", *Journal of Environmental Economics and Management* **32**(2): 251-270.
- Harrison, G.W. and J.A. List (2004), "Field Experiments", *Journal of Economic Literature* **XLII**: 1009-1055.
- Harrison, G.W. and E. Ruström (2008), "Chapter 81: Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods", *Handbook of Experimental Economics* **1**: 752-767.
- Herriges, J., C. Kling, C-C Liu, and J. Tobias (*forth.*), "What are the Consequences of Consequentiality?", *Journal of Environmental Economics and Management*.
- Johnston, R.J. (2006), "Is Hypothetical Bias Universal? Validating Contingent Valuation Responses Using a Binding Public Referendum", *Journal of Environmental Economics and Management* **52**(1):469-481.
- Kahneman, D., P. Slovic and A. Tversky (1982), *Judgment under uncertainty: heuristics and biases*. Cambridge University Press, New York.
- Kriström, B. (1990), "A Non-Parametric Approach to the Estimation of Welfare Measures in Discrete Response Valuation Studies", *Land Economics* **66**(2): 135-139.
- Kuhn, T.S. (1970), *The Structure of Scientific Revolutions. 2nd Ed.*, Chicago: University of Chicago Press.
- Landry, C.E. and J.A. List (2007), "Using *Ex Ante* Approaches to Obtain Credible Signals for Value in Contingent Markets: Evidence from the Field", *American Journal of Agricultural Economics* **89**(2): 420-429.
- List, J.A. and C.A. Gallet (2001), "What Experiential Protocol Influence Disparities Between Actual and Hypothetical Stated Values: Evidence from a Meta Analysis", *Environmental and Resource Economics* **20**(3): 251-254.
- Little, J. and R. Berrens (2004), "Explaining Disparities Between Actual and Hypothetical Stated Values: Further Investigations Using Meta Analysis", *Economics Bulletin* **3**(6): 1-13.
- Murphy, J.J., P.G. Allen, T.H. Stevens, and D. Weatherhead (2005), "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation", *Environmental and Resource Economics* **30**(3): 313-325.
- Murphy, J.J. and T.H. Stevens (2004), "Contingent Valuation, Hypothetical Bias, and Experimental Economics", *Agricultural and Resource Economics Review*, **33**(2): 182-192.
- Opaluch, J.J. and K. Segerson (1989), "Rational Roots of 'Irrational' Behavior: New Theories of Economic Decision-Making", *Northeastern Journal of Agricultural and Resource Economics Review* **18**(2): 81-95.
- Poe, G.L. J.E. Clark, D. Rondeau and W.D. Schulze (2002), "Provision Point Mechanisms and Field Validity Tests of Contingent Valuation", *Environmental and Resource Economics* **23**(1): 105-131.
- Ready, R.C., J.C. Whitehead and G.C. Blomquist (1995), "Contingent Valuation When Respondents are Ambivalent", *Journal of Environmental Economics and Management* **29**(2): 181-186.
- Satterthwaite, M. (1975), "Strategy-Proofness and Arrow Conditions: Existence and Correspondence Theorems for Voting Procedures and Welfare Functions", *Journal of Economic Theory* **10**(2): 187-217.

- Smith, V.L. (1976), "Experimental Economics: Induced Value Theory", *American Economic Review*, **66**(2): 274-279.
- Taylor, L.O., M. McKee, S.K. Laury and R.G. Cummings (2001), "Induced-Value Tests of the Referendum Voting Mechanism", *Economics Letters* **71**: 61-65.
- Vaughan, W.J. and D.J. Rodriguez (2001), "Obtaining Welfare Bounds in Discrete-Response Valuation Studies: Comment", *Land Economics* **77**(3): 457-465.
- Vossler, C.A. and M. McKee (2006), "Induced Value Tests of Contingent Valuation Elicitation Mechanisms", *Environmental and Resource Economics* **35**: 137-168.
- Vossler, C.A. and M.F. Evans (*forth.*), "Bridging the Gap Between the Field and the Lab: Environmental Goods, Policy Maker Input, and Consequentiality", *Journal of Environmental Economics and Management*.