9. Assessing the validity of stated preference data using follow-up questions

Kelley Myers, Doug MacNair, Ted Tomasi, and Jude Schneider[1]

INTRODUCTION

Stated preference (SP) studies such as contingent valuation (CV) and discrete choice experiments (DCEs) are often used to attempt measurement of willingness to pay (WTP) for environmental goods. However, concern exists that these methods do not provide data that can support valid, reliable, and meaningful WTP estimates, especially in the context of estimating non-use values for environmental goods. The foundation of all survey-based exercises is that the questions as asked by the researcher and answered by the respondent share a common understanding. This common understanding is difficult to achieve. In WTP studies, additional criteria must be met if the results are to provide data for estimating Hicksian welfare measures.[2] The criteria that must be satisfied if SP data are to be theoretically interpreted via the standard microeconomic rational choice model (RCM) have been widely discussed in the literature (e.g., Mitchell and Carson, 1989; Carson and Groves, 2007; US EPA SAB, 2009; Carson and Louviere, 2011; Bateman, 2011). General consensus exists on these criteria: that the respondents believe the information in the survey and base their responses solely on outcomes described in the survey,

[1] Respectively, Senior Economist, Cardno, Newark, DE (corresponding author, Kelley Myers@cardno.com); Technical Director, Economics and Decision Sciences, ERM, Raleigh, NC; Vice President, Cardno, Newark, DE; Senior Consultant, Cardno, Santa Barbara, CA.

[2] Economists, psychologists and others have developed more behaviorally based theories that depart from the standard microeconomic model of rational choice. While welfare measures may be developed for such theories, we focus here on the standard interpretation of rational choice and Hicksian WTP measures associated with such choice.

they treat the exercise posed in the survey as they would a real decision that affects their budget, and they answer valuation questions as rational economic agents with well-defined preferences who are trading money for economic goods.

One approach for assessing whether respondents satisfy these criteria is to use follow-up, debriefing questions. The earliest and most ubiquitous follow-up questions were "Yes/No" follow-ups based on recommendations from the National Oceanic and Atmospheric Administration (NOAA) Blue Ribbon Panel on contingent valuation. As part of a review of the use of contingent valuation to estimate lost non-use values in the context of natural resource damage assessments (NRDAs), the NOAA panel recommended the use of "Yes/No" follow-ups to determine the type of response (i.e., protest vote, yea-saying, etc.). However, the scope of follow-up questions has expanded over time (Krupnick and Adamowicz, 2006 provide a discussion). These questions may be used to "shore up the credibility of the survey" (ibid.), "to modify the estimate derived from one or more SP questions in some way" (Carson and Louviere, 2011), or to identify "problematic responses" in order to delete some responses or respondents or treat them as zeros for analysis purposes.

Despite their ubiquity, there is little consistency to either the questions posed or their use to modify analyses. First, no consensus exists on what or how many questions to ask in order to identify problematic responses. Second, most studies report results by question rather than by respondent; thus, the literature does not evaluate how many respondents had a general understanding of the tasks asked of them. Third, other than those respondents who protest the SP exercise as a whole and typically are dropped from the analysis sample, no consensus exists on what to do about problematic answers. This lack of consistency in the use of follow-up questions is troubling, as substantial proportions of respondents may give problematic answers to some of the follow-up questions and welfare estimates may be sensitive to decisions made regarding such answers.

This chapter does not solve this problem; we do not propose a theory of "problematic responses" and a practice for what to do about them. We do, however, provide some new insights into the potential prevalence of problematic responses and assess whether *respondents* are providing valid information. We focus on the pattern of responses by individual respondents to follow-up questions across a suite of debriefing questions. These questions identify whether respondents are failing to meet the criteria for satisfactory SP responses discussed above. This approach allows us to assess whether respondents "fail" on a large number of questions or only one, fail on one or many validity criteria, give responses that are correlated with observable

demographic variables, and whether validity failures are related to answers to valuation questions.[3]

The subject of our survey is valuing wetland restoration projects to reduce the effects of hypoxia in the Chesapeake Bay. The survey was Internet based and uses a sample of respondents from a web-based panel. The results show that most respondents do not meet the fundamental SP assumption that responses to valuation questions reflect carefully considered, rational economic values for the goods being evaluated in the survey. In fact, if one uses the answers to our suite of follow-up questions as a whole to identify a "core"[4] group of respondents who give unambiguously valid responses, the core would include two respondents out of a total of 1,224, both of whom were not willing to pay for environmental improvements in any of their votes. We also find that people are likely to fail more than one question within a single validity criterion. In other words, when using different types of questions to address the same topic (i.e., various types of questions and response formats that address scenario attendance), people still fail, thus reducing the likelihood that the failing response was due to response error (e.g., misinterpreting questions, marking the wrong response).[5] Further, we find little relationship between the tendency to fail the criteria and demographic variables. Hence, applying some sort of weight to the sample to match the population based on census data does not appropriately weight for the proportion of those in the population that would fail to meet the SP validity criteria. This undermines the ability to apply "econometric fixes" to problematic answers.

The rest of the chapter is organized as follows. The next section provides examples of SP studies that use debriefing questions. The third section describes our study design and data. The fourth summarizes the results, while the fifth describes the implications and paths for further research.

[3] Of course, asking a follow-up question about what a respondent was thinking when answering the primary valuation question has its methodological deficiencies. An alternative is a "think aloud" protocol in real time (e.g., Schkade and Payne, 1994) as the respondent is answering the question. However, follow-up questions are frequently used to identify "problem responses" and can trigger alternative estimators of welfare. It is this practice we address here.

[4] Bishop et al. (2011) refer to those satisfying criteria as being part of a "rational core" of respondents; here, we call those passing all questions as part of the "core."

[5] For example, we ask respondents seven different questions to assess whether they attend to the voting scenarios and outcomes described in the survey and not others. The average number of failed questions is three, and 75% of respondents fail between two and five questions.

LITERATURE REVIEW

This section discusses the use of follow-up questions in the SP literature and describes how they map to the basic principles of validity (i.e., respondents take the exercise seriously and treat it as they would a real decision, believe the information in the survey and answer valuation questions as rational economic agents with well-defined preferences). The goal is not to provide a comprehensive assessment of whether or not validity has been found to be a significant problem. Instead, we provide a description of some of the ways that it has been assessed as background information to illustrate how we developed our approach. Table 1 provides examples of the results of follow-up questions reported in the SP literature.

The most common approach to assessing whether respondents take the SP exercise seriously (i.e., view their responses as consequential) is to use questions that ask about response certainty. Using this approach, respondents are asked how certain they are that they would actually pay the amount, or vote as they indicated they would in the survey. Scenario acceptance, or belief in the information provided, requires that respondents value the good described in the survey (and not some other good of their own construction) based on the stated price (and not some other price they believe they would pay). A number of studies ask follow-up questions to test whether respondents believe the survey scenario (Carson et al., 1994, 2003; Krupnick et al., 2002; Banzhaf et al., 2006, 2011; Bishop et al., 2011). These studies ask whether the individual believed the outcomes described would occur, if they believed they would have to pay the amount shown, or if they valued something larger than the good in question.

Finally, the third criterion requires that respondents exhibit utility maximizing behavior and make trade-offs according to standard compensatory methods. Examples of behaviors that violate this criterion include problematic attitudes such as yea-saying or purchasing moral satisfaction, protest responses, using simplifying decision heuristics rather than careful evaluations, and ignoring certain attributes of the SP question. Follow up questions are often used to identify these types of behaviors and adjust the WTP values accordingly.

Our literature review yields three insights that guided our study design. First, because of the widespread use of follow-up questions in CV surveys, we expected that almost all recent DCE studies would use follow-up questions to test validity comprehensively. However, the proportion of DCE studies using follow-up questions to test validity is smaller than we expected,[6] and most studies that do use follow-up questions only focus

[6] In their review of supporting questions in DCEs, Krupnick and Adamowicz (2006)

Table 1 Summary of SP studies that use debriefing questions

Author(s) and Year	Good	Question Topic(s)	Percentage of Problematic Responses[a]
Contingent valuation studies			
Li and Mattson (1995)	Forests	Response certainty	64
Champ et al. (1997)	Open space	Response certainty	Not reported
Champ and Bishop (2001)	Wind-generated electricity	Response certainty	48
Poe et al. (2002)	Green electricity	Response certainty	21
Banzhaf et al. (2006)	Ecosystem services	"Yea-saying" and protest no's	59
Carson et al. (2003)	Damages from oil spill	Scenario acceptance	Not reported
Carson et al. (1994)	Damages from DDT and PCBs	Scenario acceptance	48
Krupnick et al. (2002)	Mortality risk	Scenario acceptance	40
Discrete choice experiments			
Olsson (2005)	Cod	Response certainty	71
Ready et al. (2010)	Wild animals	Response certainty	Mean certainty 6.5 reported
Bishop et al. (2011)	Hawaiian coral reefs	Scenario acceptance	54
Scarpa et al. (2009)	Alpine grazing areas	Ignoring attributes	40–80
Carlsson et al. (2010)	Environmental quality	Ignoring attributes	54
Banzhaf et al. (2011)	Ecosystem services	"Yea-saying" or hypothetical bias	9[b]
Cameron et al. (2010)	Major illness/ injury	Scenario replacement/ adjustment	Not reported
Kataria et al. (2012)	Water quality	Scenario acceptance	64

Notes:
a. Represents the highest number reported from all questions.
b. Only report frequency of responses to one out of 34 questions.

on one test of validity. Establishing the validity of a stated choice survey is fundamental, but assessing validity does not appear to be a standard practice in DCE studies. Second, numerous studies report results that show significant portions of the population giving a problematic response to the follow-up question, casting doubt on the DCE's validity. Third, these studies tend to report sample proportions answering a question in an invalid fashion for each question separately. The pattern of responses across questions and across question topics by an individual respondent is not investigated.

STUDY DESIGN AND DATA

Our study developed follow-up questions as part of a stated choice survey about reducing hypoxia in Chesapeake Bay. Survey development occurred between August 2010 and September 2011 with the aid of two focus groups and four one-on-one interview sessions. The survey has four sections, which is consistent with current practices in DCEs (see Bateman et al., 2002 for more information).

The first section introduces respondents to Chesapeake Bay and describes the causes and impacts of hypoxia and how restoring coastal wetlands can reduce these effects. The first section also asks some general warm-up questions about environmental attitudes. The second section describes a potential program for reducing hypoxia in Chesapeake Bay by restoring coastal wetlands. This section includes a description of the policy change, the institution for providing this change, and the payment mechanism. In our survey, the policy change is a second phase of restoration to build on restoration that has already occurred in Phase 1 (thus mitigating the desire to vote yes to "do something" for the environment, since something already has been done). If approved, Phase 2 would require a one-time payment through increased income taxes for all US households. We select a national income tax as the payment mechanism since the benefits of the restoration are not limited by geographic location.[7] The pages that follow describe the attributes affected by the program, which include acres of restored wetlands, bird diversity, days without excess algae, fish and shellfish abundance, public access to wetlands, and a Chesapeake Bay

state that "[a] surprisingly large number of stated choice surveys do not use debriefing questions. . .that ask respondents what they felt or thought as they read text or answered questions." Our review also supports this finding.

[7] Using a general tax as a payment mechanism is one of two types of coercive payment mechanisms commonly used in DCEs (see Carson and Louviere, 2011 for a discussion).

ecosystem health score.[8] The attributes were developed over the course of a year through a combination of consultation with ecologists, subject matter experts, and focus group respondents.

We also designed several of the ecological attributes, including the Chesapeake Bay ecosystem health score, by following guidelines for ecological indicators in SP valuation developed by Johnston and collaborators (Johnston et al., 2011, 2012).

The third section describes the voting format and includes a reminder about some of the pros and cons of voting for a restoration program. The pros include belief that reducing hypoxia in the Chesapeake Bay is worth the cost and is a good use of tax dollars, and that the cost of the tax increase is within a respondent's budget. The reasons to vote against the program include belief that it is not worth the cost, not a good use of tax dollars, and not within the respondent's budget. After a sample vote, each respondent votes on five different combinations of restoration outcomes. In each vote, respondents have the option to choose the status quo (keep the amount of restoration completed in Phase 1 and pay nothing) or to choose one of two alternative restoration programs at an additional cost to their household. To generate the choice sets, we used SAS market research macros to generate a D-Optimal fractional factorial design out of the full factorial ($4^6 * 2^1 = 8{,}192$ alternatives). This generation produced 24 choice pairs that we blocked into six groups of four.[9] A sample choice set is shown in Figure 1.

The fourth and final section of the survey contains the debriefing questions, followed by standard socioeconomic and demographic questions.

Our data come from 1,224 respondents enrolled in a web-enabled panel maintained by Research Now.[10] Of the sampled respondents, 27% said they had visited Chesapeake Bay. The average income of our sample was generally in line with 2010 census data, but the income range from $25,000 to $74,900 was slightly over-represented and the higher ranges were

[8] For a list of the attributes, their descriptions, and their respective levels, please contact the corresponding author.

[9] Each respondent saw four choice pairs that came from the experimental design plus a fifth pair that was common across all respondents. The decision to use five choice sets was based on a review of environmental studies that use DCEs to value similar goods. For example, Carlsson et al. (2003) present respondents with four choice sets each, while Hoehn et al. (2005) use five. Although it is not uncommon to use more choice sets (i.e., Birol et al., 2006 use eight), some empirical evidence suggests that cognitive burden increases with the number of choice sets (Bech et al., 2011).

[10] Research Now uses a "by-invitation only" methodology to member selection by partnering with a variety of businesses. Invitees already have a pre-existing relationship with the company that invited them, guaranteeing a high-quality panel while minimizing duplication, fraudulent responses, and professional survey takers.

Outcomes	Current Situation Without Phase 2	Design A	Design B
Wetland acres	52% of goal 13,000 acres out of 25,000 acres (0 more acres)	60% of goal 15,000 acres out of 25,000 acres (2,000 more acres)	90% of goal 22,750 acres out of 25,000 acres (9,250 more acres)
Bird diversity	30 species	30 species	36 species
Days without excess algae	20% of summer days 20 out of 100 days	20% of summer days 20 out of 100 days	80% of summer days 80 out of 100 days
Public access to wetlands	Some areas have access	Additional access	No additional access
Fish and shellfish abundance score	59 out of 100	59 out of 100	76 out of 100
Chesapeake Bay ecosystem function score	46 out of 100	46 out of 100	55 out of 100
Total one-time cost per US household	$0	$50	$295
Please make your selection (Choose one):	○	○	○

Figure 1 Example valuation question

slightly under-represented; 51% were males compared to 49.2% from the census data.[11]

METHODS AND RESULTS

Our analysis has three basic components. First, we review the frequency distributions of responses to each of the debriefing questions and impose three degrees of rigor for specifying whether a respondent satisfies the validity criteria.[12] Second, we examine each respondent's answers to see if

[11] Our sample was drawn from a nationwide population, stratified by 30% from the Mid-Atlantic region, 30% from the West, and 40% from the rest of the country.

[12] The frequency distributions are available upon request from the corresponding author.

a pattern exists by respondent (across the follow-up questions). Third, we use multivariate regression analysis to investigate whether the tendency to meet the validity criteria is associated with a particular type of response to the voting question (i.e., all yes votes, etc.), and to see if respondent characteristics can predict whether a respondent is more or less likely to pass the validity criteria.

Descriptive Statistics

Our analyses are based on a portion of the full sample (960 out of 1,224) as we exclude protest no's (i.e., people who stated they did not trust the government, did not believe in tax increases of any kind, or did not feel they should have to pay for the good). Based on our focus groups and one-on-ones, we use various types of questions and response formats, which include multiple choice, open-ended, and contrasting statements (a question that asks respondents to indicate whether they agree with contrasting statements on either end of a five-point scale) and use several different question types to address the same topic (i.e., attending to the scenario, believing responses will affect the outcome, etc.).

Clearly our response options give some leeway in determining what constitutes a "problem" in a response. For some of the questions, we construct three classes of "rigor" for specifying when a respondent satisfies the validity criteria: *stringent*, *average*, and *lenient*.[13] The stringent approach leads to the smallest number of respondents meeting the validity criteria: respondents meet the criteria only if their answers are unambiguously valid. These respondents are most clearly part of the "core" of respondents. For example, respondents were asked several questions with a contrasting statement response format where an unambiguously valid response was to the far left (response option A) and an invalid response was on the far right (response option E), with five total response categories from A to E. In the stringent approach, respondents who chose A or B are regarded as satisfying the validity criterion for this type of question. The lenient approach accepts more response categories as meeting the criteria and so expands the size of the core. The average approach is in the middle of these two.

In general, for questions that have only two response options (one that is unambiguously valid and one that is not), we do not specify a degree of rigor. A respondent either gives a valid response, or does not. Also, we do not specify a degree of rigor if a question is only shown to a subset of the

[13] A detailed description about what constitutes a stringent, average, and lenient classification for each question is available upon request from the corresponding author.

entire sample (i.e., a follow-up question based on a response to a previous question). In both cases, this is illustrated in Table 2 where the frequency of response is the same across all categories (lenient, average, and stringent).

Table 2 provides a complete list of valid responses to each of the questions and the percentage meeting the validity criterion for each question. Looking at some of the responses that did not vary by the degree of rigor, Table 2 shows that 20% of the sample gives a valid response to a question regarding how they considered costs of the restoration program when making a choice. A valid response to this question includes "I thought only about how I and/or my family would be affected by the cost," whereas an invalid response includes "I thought about an amount that would be fair for most people to pay" and "I thought about an amount that would get a lot of people to vote yes." Additionally, less than 40% valued the program as described (i.e., did not consider health effects when deciding their votes, which the survey explicitly excluded as a benefit of the program), thought they would have to pay the amount shown, did not include other outcomes like reducing toxic chemicals not part of the scenario, and did not consider that voting for the program would increase the chances of the government starting a similar program near them.

Using the stringent approach to identifying valid responses, 28% of respondents saw the results as consequential (i.e., thought that the survey responses would be used to decide whether taxes would be collected). Twenty-one percent of the sample thought program outcomes should be chosen based on people's answers to questions in surveys like this one and 25% thought that survey sponsors want to find out how much the public values the program.

Cumulative Assessment of the Validity of Responses

Next, we examine responses to the follow-up questions at the respondent level. Figure 2 provides the cumulative percentage of the respondents who give invalid responses by degree of rigor. Using the most lenient assessment, 50% of the respondents failed at least six questions. Using the most stringent assessment, 50% of the respondents fail at least nine questions.

We next identify respondents who provide valid responses to all questions and, therefore, make up the "core" of respondents. Table 3 shows that only two people are in the core using the lenient approach to inclusion, one person is in the core using the average approach, and the stringent core is empty. Moreover, the two people who do remain in the lenient core (and therefore clearly can be judged to engage in the survey as real, understand the choices being asked of them, and respond in accord with economic

Table 2 Percentage of sample that gives a valid response by question (n = 960)

Valid Response	Percentage of Sample		
	Stringent	Average	Lenient
Thought only about how I and/or my family would be affected by the cost	20	20	20
I think the survey responses will be used to decide if taxes will be collected for the program	28	56	81
My votes will affect the size and scope of the program	35	69	87
When I voted for a design that costs money it was to show support for the program and I am willing to pay the tax	55	79	93
Thought about things that I would not be able to buy for my family or about the other causes that I would not be able to support in order to pay for the program	62	62	62
I would vote the same way if the program were actually on the ballot (Certainty > 50%, > 30%, > 0%)	71	91	97
I did NOT consider that the program would protect the health of the people who eat fish from the Bay when deciding my votes	34	34	34
If this design is implemented, I think I would end up actually paying the amount shown	35	35	35
When I voted, I thought the wetland restoration program would only have a significant effect on reducing the excess phosphorus and nitrogen that causes hypoxia	36	36	36
I did NOT consider that if enough people voted for the program, it would increase the chances that the government would start a program to restore wetlands near me when deciding my votes	40	40	40
I hope and believe that the tax money spent will only be spent on the program	41	41	41
I hope and believe that the program will provide the restoration outcomes as described in the survey	62	62	62
I hope and believe that the program will reduce hypoxia in Chesapeake Bay as described in the survey	71	71	71
If the designs for the programs described here all cost the same amount and were funded by existing sources they should be chosen based on people's answers to questions in surveys like this one	21	51	75

Table 2 (continued)

Valid Response	Percentage of Sample		
	Stringent	Average	Lenient
Sponsors of the survey don't know whether to fund the program and want to find out how much the public values this program	25	55	79
When I made my choices in each of the votes I found the choices straightforward and carefully compared all of the outcomes as described in the survey	43	58	58

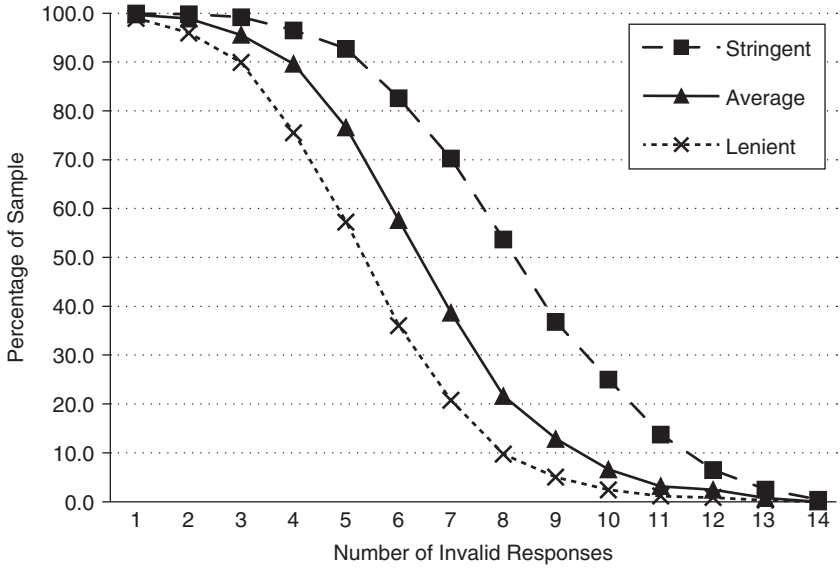

Figure 2 Cumulative percentage of sample that fails at least one question

rationality) voted against contributing taxes for environmental improvements in all five votes.

These results demonstrate that substantial proportions of the respondents do not provide responses to the follow-up questions that comport with the validity criteria. Of course, asking a large set of questions increases the chances that a respondent gives at least one invalid response. Hence, we do not necessarily propose that those not in the core be dropped from the

Table 3 Number of respondents in core

Vote Type	Lenient	Average	Stringent
Voted yes at least once	0	0	0
Voted all no	2	1	0
Total	2	1	0

analysis sample.[14] However, we find the portion of responses indicating a failure to meet validity criteria very troublesome.

Regression Analysis

This section examines the relationship between the follow-up questions and how people respond to the voting scenarios. It also explores whether respondent characteristics are reasonable predictors as to whether people will give valid or invalid responses to the follow-up questions. Table 4 shows the results of a zero-inflated Poisson model that regresses the responses to the follow-up questions for each individual on the number of status quo votes. We use the average approach to identifying responses as valid, and code each variable so that a 1 equals a valid response, 0 otherwise. The top half of the table indicates whether the response to the follow-up question has an influence on the total number of status quo votes, whereas the bottom half of the table is a logit regression that indicates whether a respondent is more likely to be a certain zero (i.e., "all yes" voter) based on their response to the question. The results show that seven questions influence the probability of being a certain zero (i.e., someone who votes yes to all five votes). The results also show that a subset of these questions influences the number of no votes.

For example, giving the following valid responses lowers the probability of voting yes to the restoration program in all five votes: considering *only* how their family would be affected by the taxes when voting (*Family_only*), and not how other people would be affected, that the votes will affect the size and scope of the restoration program (*Vote_affectscope*), and that the survey responses will be used to decide if taxes will be collected for the program (*Vote_affecttaxes*). In other words, respondents who give valid responses to these questions are less likely to be "yes" voters. Table 4 also indicates that valid responses to *Family_only* and *Vote_affecttaxes*

[14] With potential for response errors, as the number of questions increases, the probability of answering all correctly goes to zero even if the respondent is in the core. This outcome begs the question of what to do with respondents that fail some significant fraction of follow-up questions assessing validity, which is beyond the scope of this chapter.

Table 4 Zero-inflated Poisson regression by valid responses on number of "no" votes

Variable[a]	Coefficient	t-Statistic
NP_only	−0.01	−0.09
Family_only	0.16	2.53**
Didnotconsider_chances	0.05	0.68
Didnotconsider_health	0.02	0.26
40%_certain	−0.15	−1.94*
Hope_programworks	−0.13	−1.64
Hope_moneyspent	0.01	0.78
Hope_outcome	0.06	0.19
People_choose	0.02	0.38
Sponsors_dontknow	0.03	0.51
Vote_affecttaxes	0.05	0.76
Vote_affectscope	0.11	1.76
Willingtopay_tax	−0.41	−4.79**
No_decisionstrategy	0.07	1.2
Believe_payamtshown	−0.25	−2.33*
Constant	1.10	9.58
Inflate (Logit regression)		
NP_only	−0.11	−0.62
Family_only	−0.82	−3.76**
Didnotconsider_chances	−0.40	−2.17*
Didnotconsider_health	−0.33	−1.73
40%_certain	1.47	3.56**
Hope_programworks	0.51	2.17*
Hope_moneyspent	−0.06	−0.3
Hope_outcome	0.17	0.89
People_choose	−0.23	−1.35
Sponsors_dontknow	−0.05	−0.28
Vote_affecttaxes	−0.37	−2.17*
Vote_affectscope	−0.51	−2.85**
Willingtopay_tax	1.21	6.49**
No_decisionstrategy	−0.01	−0.08
Believe_payamtshown	0.02	0.10
Constant	−1.31	−2.83

Notes:
* Indicates significance at the 95% level of confidence; ** indicates significance at the 99% level of confidence.
a. A description of each variable is available upon request from the corresponding author.

increases the number of "no" votes. The overall conclusion from this analysis is that it takes more than just a single follow-up question or a single type of question (i.e., response certainty, etc.) to evaluate how people respond to the voting question. Our results indicate that many of our questions either affect how people responded to the vote (i.e., were all yes voters) or affect the number of times a person chooses the status quo.

In a review of alternative methods of valuing environmental goods and services, US EPA SAB (2009) stated that a key criterion for choosing an approach was whether or not the method provides a reliable way to extrapolate from the respondents to the target population. The regression analysis below provides evidence that a reliable extrapolation approach may not exist. First, one needs to be reasonably certain that the respondents are a true random sample of the population. However, if people who fail to satisfy validity criteria are more likely to vote for a tax for an environmental program, it is reasonable to believe they may also be more likely to respond to the survey. Therefore, randomness cannot be assumed. However, if the propensity to pass validity tests is closely linked to demographics, then sampling weights could be used to adjust the results. Unfortunately, no strong link exists between demographics and propensity to satisfy the criteria and, thus, no such simple weighting scheme to match the sample to the population is available.

To explore this, we use a binary logit model to regress demographics (age, income, and gender) on the response to each of the follow-up questions listed in Table 4. All of the adjusted R^2s are low and in most cases, an F-test indicates that the coefficients on at least two of the variables (age, income or gender) are not significantly different from zero in each of the regressions. However, this is not consistent across questions, making it difficult to identify a consistent pattern of respondent characteristics that explains responses to any of the follow-up questions. DCE studies have used a wide variety of techniques to make "adjustments" for respondents who don't appear to be providing valid responses. Reviewing and evaluating those approaches is beyond the scope of this chapter. Our point here is that the propensity to satisfy validity criteria may be so idiosyncratic that no reliable method may exist for determining the appropriate percentage of results to adjust to extrapolate to the population.

DISCUSSION AND CONCLUSIONS

General consensus exists in the literature that respondents to SP studies should attend to the scenarios and outcomes described in the survey and not others, take the exercise seriously and treat it as they would a

real decision that affects their budget, and answer valuation questions as rational economic agents with well-defined preferences who are trading money for specific economic goods. In our survey, a large portion of our sample fails to meet these criteria. In fact, when examining all of each respondent's responses to the entire suite of follow-up questions, our sample yielded no more than two respondents out of 1,224 who answered the follow-up questions in a manner consistent with meeting the validity criteria; the average respondent failed to give a valid response, on average, to six or nine questions depending on degree of rigor in coding responses.

The majority of respondents (more than 50%) in our survey valued something other than reducing the effects of hypoxia, considered other elements of cost than how their family would be affected by the cost of the program, did not believe they would have to pay the amount shown in the vote, and/or thought that voting for the program would increase the chances of starting a similar program near them. We also find that people who vote yes for a program at least once are less likely to give a valid response and both of the respondents in the core that do give valid responses voted not to pay for the environmental program in all five votes.

Should policy decisions and legal damages be assessed using information obtained from people who appear to give invalid responses to follow-up questions such as these? What should be done with the results of answers to follow-up questions such as those we obtained? We do not propose answers to these questions, but our analysis suggests the questions are important.

It has been argued that when "state of the art" survey design and administration is employed, the results from SP studies can represent the population's true monetary values in an unbiased fashion (Ryan and Spash, 2011). However, existing literature and the results of this study may show that the inconsistent and invalid responses may be more endemic to SP methods and potentially resistant to changes in survey designs.

REFERENCES

Banzhaf, H.S., D. Burtraw, D.A. Evans, and A. Krupnick (2006), "Valuation of natural resource improvements in the Adirondacks," *Land Economics*, **82**(3), 445–64.

Banzhaf, H.S., D. Burtraw, S. Chung, D.A. Evans, A. Krupnick, and J. Siikamaki (2011), "Valuation of ecosystem services in the southern Appalachian Mountains," paper presented at the Annual Meeting of the Association of Environmental and Resource Economics, Seattle, WA.

Bateman, I.J. (2011), "Valid value estimates and value estimate validation: Better methods and better testing for stated preference research," in J. Bennett (ed.), *The*

International Handbook on Non-Market Environmental Valuation, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 322–52.
Bateman, I.J., R.T. Carson, B. Day, M. Hanemann, N. Hanley, and T. Hett et al. (2002), *Economic Valuation with Stated Preference Surveys: A Manual*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
Bech, M., T. Kjaer, and J. Lauridsen (2011), "Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment," *Health Economics*, **20**(3), 273–83.
Birol, E., K. Karousakis, and P. Koundouri (2006), "Using a choice experiment to account for preference heterogeneity in wetland attributes: The case of Cheimaditida wetland in Greece," *Ecological Economics*, **60**(1), 145–56.
Bishop, R.C., D.J. Chapman, B. Kanninen, J.A. Krosnick, B. Leeworthy, and N.F. Meade (2011), *Total Economic Value for Protecting and Restoring Hawaiian Coral Reef Ecosystems: Final Report*, Silver Spring, MD: NOAA Office of National Marine Sanctuaries, Office of Response and Restoration, and Coral Reef Conservation Program.
Cameron, T.A., J.R. DeShazo, and E.H. Johnson (2010), "Scenario adjustment in stated preference research," *Journal of Choice Modelling*, **4**(1), 9–43.
Carlsson, F., P. Frykblom, and C. Liljenstolpe (2003), "Valuing wetland attributes: An application of choice experiments," *Ecological Economics*, **47**(1), 95–103.
Carlsson, F., M. Kataria, and E. Lampi (2010), "Dealing with ignored attributes in choice experiments," *Environmental and Resource Economics*, **47**(1), 65–89.
Carson, R. and T. Groves (2007), "Incentive and informational properties of preference questions," *Environmental and Resource Economics*, **37**(1), 181–210.
Carson, R. and J. Louviere (2011), "A common nomenclature for stated preference elicitation approaches," *Environmental and Resource Economics*, **49**(4), 539–59.
Carson, R., M. Hanemann, R.J. Kopp, J.A. Krosnick, R. Mitchell, and R. Presser et al. (1994), *Prospective Interim Lost Use Value Due to DDT and PCB Contamination in the Southern California Bight: Volume II*, La Jolla, CA: US Department of Commerce (NOAA).
Carson, R., R. Mitchell, M. Hanemann, R.J. Kopp, S. Presser, and P.A. Ruud (2003), "Contingent valuation and lost passive use. Damages from the Exxon Valdez oil spill," *Environmental and Resource Economics*, **25**(3), 257–86.
Champ, P. and R. Bishop (2001), "Donation payment mechanisms and contingent valuation: An empirical study of hypothetical bias," *Environmental and Resource Economics*, **19**(4), 383–402.
Champ, P., R. Bishop, T. Brown, and D. McCollum (1997), "Using donation mechanisms to value nonuse benefit from public goods," *Journal of Environmental Economics and Management*, **33**(2), 151–62.
Hoehn, J., F. Lupi, and M. Kaplowitz (2010), "Stated choice experiments with complex ecosystem changes: The effect of information formats on estimated variances and choice parameters," *Journal of Agricultural and Resource Economics*, **35**(3), 568–90.
Johnston, R.J., E.T. Schultz, K. Segerson, and E.Y. Besedin (2011), "Bioindicator-based stated preference valuation for aquatic habitat and ecosystem service restoration," in J. Bennett (ed.), *The International Handbook on Non-Market Environmental Valuation*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
Johnston, R.J., E.T. Schultz, K. Segerson, and E.Y. Besedin (2012), "Enhancing

the content validity of stated preference valuation: The structure and function of ecological indicators," *Land Economics*, **88**(1), 102–20.
Kataria, M., I.J. Bateman, T. Christensen, A. Dubgaard, B. Hasler, and S. Hime et al. (2012), "Scenario realism and welfare estimates in choice experiments – a non-market valuation study on the European water framework directive," *Journal of Environmental Economics and Management*, **94**(1), 25–33.
Krupnick, A. and V. Adamowicz (2006), "Supporting questions in stated choice studies," in B. Kanninen (ed.), *Valuing Environmental Amenities Using Stated Choice Studies*, Dordrecht: Springer.
Krupnick, A., A. Alberini, M. Cropper, N. Simon, B. O'Brien, and R. Goeree et al. (2002), "Age, health and willingness to pay for mortality risk reductions: A contingent valuation survey of Ontario residents," *Journal of Risk and Uncertainty*, **24**(2), 161–86.
Li, C.Z. and L. Mattsson (1995), "Discrete choice under preference uncertainty: An improved structural model for contingent valuation," *Journal of Environmental Economics and Management*, **28**(2), 256–69.
Mitchell, R. and R. Carson (1989), *Using Surveys to Value Public Goods: The Contingent Valuation Method*, Washington, DC: Resources for the Future.
Olsson, B. (2005), "Accounting for response uncertainty in stated preference methods," paper presented at the EAERE Congress, Bremen, Germany.
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), 105–31.
Ready, R.C., P. Champ, and J. Lawton (2010), "Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment," *Land Economics*, **86**(2), 363–81.
Ryan, A. and C. Spash (2011), "Is WTP an attitudinal measure? Empirical analysis of the psychological explanation for contingent values," *Journal of Economic Psychology*, **32**(5), 674–87.
Scarpa, R., T. Gilbride, D. Campbell, and D.A. Hensher (2009), "Modelling attribute non-attendance in choice experiments for rural landscape valuation," *European Review of Agricultural Economics*, **36**(2), 151–74.
Schkade, D. and J. Payne (1994), "How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation," *Journal of Environmental Economics and Management*, **26**(1), 88–109.
United States Environmental Protection Agency (EPA) Scientific Advisory Board (SAB) (2009), *Valuing the Protection of Ecological Systems and Services*, Washington, DC: EPA.