An Adding-Up Test on Contingent Valuations of River and Lake Quality

by

William Desvousges[a*], Kristy Mathews[b], Kenneth Train[c]

Forthcoming, *Land Economics*

[a]President, W.H. Desvousges & Associates, P.O. Box 99203, Raleigh, NC 27624, william.desvousges@whdesvousgesassociates.com

[b]Owner and Principal,104 McWaine Lane, Cary, NC 27513, kmathews1@nc.rr.com

[c]Adjunct Professor, Department of Economics, University of California, Berkeley, 530 Evans Hall #3880, Berkeley, CA 94720-3880, train@econ.berkeley.edu.

[*] Corresponding author, 919-413-6225. The research was supported in part by BP. The analysis and conclusions in this paper reflect the views of the authors and should not be attributed to anyone else.

An Incremental Adding-Up Test on Contingent Valuations of River and Lake Quality

by

William Desvousges, Kristy Mathews, and Kenneth Train

Abstract

This paper tests for a contingent valuation (CV) study of a bundle of water quality improvements, whether the sum of the estimated willingness to pay (WTP) for each individual part of the package, evaluated incrementally, equals the estimated WTP for the entire bundle – as implied by standard utility theory. This is the first application of the adding-up test using incremental parts and Carson and Groves' (2007) procedures to induce truthful respondents' answers. The test is failed, which implies that either the CV method did not elicit truthful answers or that respondents' true preferences are inconsistent with standard utility theory.

1. Introduction

Contingent valuation (CV) is a survey procedure designed to estimate respondents' willingness to pay (WTP) for natural resource services. See Carson and Hanemann (2005) for a review. One of the most prominent concerns about CV is whether the estimated WTP from CV studies varies adequately with the amount, extent, or, more generally, "scope," of the environmental good.[i] This concern was emphasized by a panel of experts that the National Oceanic and Atmospheric Administration (NOAA) convened with the purpose of making recommendations about the reliability of CV. The panel concluded that they would judge the findings of a CV study to be unreliable if it evidenced "Inadequate responsiveness to the scope of the environmental insult," and said that the burden of proof "must rest" with the researchers who designed and implemented the study (Arrow et al. 1993).

Researchers have implemented scope tests that examine whether the estimated WTP in CV studies increases (or at least does not decrease) when the scope of environmental benefits is expanded. See Carson (1997) and Desvousges, Mathews, and Train (2012) for reviews. However, passing the scope test (i.e., finding that estimated WTP increases with scope) does not imply that the estimated response is adequate in magnitude. Members of the expert panel (Arrow et al. 1994) explicitly stated that the scope test does not address their concern about adequacy of response to scope.[ii]

In this paper, we discuss and implement Diamond et al.'s (1993) adding-up test, which has important implications for the issue of adequate response. Diamond et al. point out that standard utility theory implies a relation called the "adding-up condition," namely, that the WTP for one good, plus the WTP for a second good once the consumer has paid for and obtained the first one, is necessarily equal to the consumer's WTP for both goods combined. A more precise definition is given in section 2 below; colloquially, the condition is often expressed as "the whole equals the sum of the incremental parts," with "incremental" meaning that the second good is evaluated after having paid for and received the first good.

Diamond et al.'s test of the adding-up condition is implemented as follows: one group of respondents is asked their WTP for one good; a second group is told that this good has already been provided to them and asked their WTP for a second good; then a third group is asked their WTP for both goods. If the WTP from the first group plus the WTP from the second group equals the WTP from the third group, then the adding-up test is passed: the WTPs are consistent with the adding-up condition.

The adding-up test can address the NOAA panel's concern about adequate response to differences in scope. If the sum of WTPs for individual benefits, evaluated incrementally, equals the WTP for all of them combined (i.e., the adding-up test is passed), then the response to differences in scope is clearly adequate. However, if the sum of the estimated WTPs for the incremental benefits exceeds the estimated WTP for all of them combined (i.e., the adding-up test is failed), then questions arise about the adequacy of the CV responses to changes in scope.

Despite the potential value of the test, no studies since Diamond et al. have applied an adding-up test to incremental parts of public goods.[iii] Several studies have examined adding-up for non-incremental benefits.[iv] In particular, they elicited consumers' WTP for one good, the WTP for a second good, and the WTP for the two goods combined; however, the WTP for the second good was evaluated without the consumer having obtained the first good. The adding-up condition applies only for goods that are obtained incrementally. As noted by the authors of these studies, as well as others, failure of adding-up on non-incremental parts can arise because of diminishing marginal utility or substitution, both of which are consistent with standard utility theory. Diamond et al. specified the test for incremental benefits, such that diminishing marginal utility and substitution, to the extent they exist, are incorporated into the valuations.

The previous CV studies on adding-up are potentially problematic for another reason as well. Carson and Groves (2007) identified features of the CV scenario that are designed to induce truthful answers from respondents, a concept that they call consequentiality. Except for de Zoysa (1995),[v] all of the previous studies that examined the adding-up condition have used CV methods that differ from those designed to induce truthfulness.

Failures of the adding-up test in these studies could therefore be attributed to the lack of consequentiality in the design of the CV scenarios.

In this paper, we test the adding-up condition using incremental parts on CV scenarios that are designed to induce truthful answers. To our knowledge, this is the first investigation that satisfies both these criteria. We implement the test on a study by Chapman et al. (2009) that evaluated a restoration program for a specified river system and lake in Oklahoma. The lake and river system had been polluted by "chicken litter" that caused an overgrowth of algae; the study estimated the WTP for a program to put alum on the water to reduce the algae. We chose this study because it represents the current state-of-the-art for CV and its scenarios were designed to meet the consequentiality conditions identified by Carson and Groves (2007). Also, the study had already described its program in incremental parts for the purposes of a scope test and had developed the survey instrument for one of these increments (in addition to the instrument for the program as a whole), with wording that described to respondents how the first increment had already been funded and provided. This feature allowed us to implement an adding-up test with minimal changes in the questionnaires.

We find that the adding-up condition does not hold in this study, with the sum of the WTP for the incremental parts being three times greater than that for the whole. This result implies that either (i) the CV procedure, with incremental parts and procedures designed for truthfulness, did not elicit the true preferences of consumers, or (ii) consumers' true preferences are not consistent with standard utility theory.

The remainder of the paper is organized as follows. Section 2 describes the adding-up condition more formally, with Section 3 enumerating practical issues that need to be

considered in implementing the adding-up test. Section 4 describes past studies whose designs allowed an adding-up test on incremental parts, even if the test were not performed. Section 5 discusses the study by Chapman et al. (2009) as it relates to the adding-up test, and, in Section 6, we describe the increments that we specified for the test and the way that the original surveys were revised for the additional increments. Results are given in Section 7. Income effects are investigated in Section 8. The paper concludes with a discussion of the interpretation and implications of the results.

2. The Adding-Up Test

Diamond et al. (1993) explain the adding-up test through analogy:

> For instance, consider asking one group of people how much they are willing to pay for a cup of coffee. Ask a second group how much they would be willing to pay just for a doughnut if they already had been given a cup of coffee. Ask a third group how much they would be willing to pay for a cup of coffee and a doughnut. The value obtained from the third group should equal the sum of the values obtained from the first two groups if the answers people give reflect underlying economic preferences (p. 48).

More formally, let $e(x, y, p, u)$ be the consumer's expenditure function at prices *p* for private goods, utility level *u*, and levels *x* and *y* of two public goods. Consider a program that increases the quantity of the public goods from x_0 to x_1 and y_0 to y_1. WTP for this improvement is defined as $WTP(x_1, y_1|x_0, y_0) \equiv e(x_0, y_0, p, u) - e(x_1, y_1, p, u)$. Adding and subtracting terms gives:

$$\begin{aligned} WTP(x_1, y_1|x_0, y_0) &= e(x_0, y_0, p, u) - e(x_1, y_0, p, u) + e(x_1, y_0, p, u) - e(x_1, y_1, p, u) \\ &\equiv WTP(x_1, y_0|x_0, y_0) + WTP(x_1, y_1|x_1, y_0) \end{aligned}$$

which is the adding-up condition. The same relation occurs for a program that increases x_0 to x_1 without changing *y*, with the increments defined by an intermediate level x_i with $x_0 < x_i < x_1$. Note that the only assumptions that are required for the adding-up condition are those required for the existence of the expenditure function.

The adding-up test extends the scope test in an informative way. For the standard scope test, one group of respondents is asked about their WTP for a specified set of benefits, $WTP(x_1, y_1|x_0, y_0)$, and a second group is asked about their WTP for a subset of these benefits, e.g. $WTP(x_1, y_0|x_0, y_0)$. The adding-up test is implemented by also asking another group of respondents about their WTP for the benefits included in the first set but excluded from the second set, with the benefits defined incrementally, e.g., $WTP(x_1, y_1|x_1, y_0)$.

This extension resolves the uncertainties that arise in interpreting scope test results. Suppose, for example, that the scope test is passed when comparing $WTP(x_1, y_1|x_0, y_0)$, with $WTP(x_1, y_0|x_0, y_0)$. As stated above, this result does not imply that the magnitude of the estimated difference is adequate. The adding-up test provides a means to evaluate the magnitude of this difference, by testing whether it equals the directly estimated $WTP(x_1, y_1|x_1, y_0)$. Suppose instead that the scope test fails. As stated above, diminishing marginal utility and substitution can cause little or no response, which can lead to failure of the scope test. The adding-up test assesses whether the failure reflects these kinds of preferences, by determining whether $WTP(x_1, y_1|x_1, y_0)$ is sufficiently small.[vi]

Haab et al. (2013) state that the adding-up test imposes additional structure on preferences beyond that imposed by the scope test and that the additional structure is unnecessary. For the scope test, they say that "A simple theoretical model of WTP, a difference in expenditure functions with changes in quality or quantity, can be used to show that WTP is nondecreasing in quality or quantity (Whitehead, Haab, and Huang

1998).” The same theoretical model, with differences in expenditure functions (as described above), is all that is needed to show the adding-up condition. The assumptions that Whitehead, Haab, and Huang (1998) use to show non-negative scope effects for the scope test are sufficient to show the adding-up condition for the adding-up test. No additional assumptions or structure is required.[vii]

The adding-up test examines a different implication of utility theory than the scope test, which might explain Haab et al.’s concern. However, the difference constitutes one of the potential values of the test: the adding-up test can address the issue that NOAA’s expert panel enumerated, while the scope test does not. Given that the panel said that the burden of proof “must rest” on the researcher, the adding-up test seems particularly useful.

More generally, the adding-up test can be considered similar to the research on the WTP and willingness to accept (WTA) discrepancy. Haab et al. (2013) note that evidence of a WTP/WTA disparity is “a call for the curious researcher to more closely examine the assumptions and structures leading to these seemingly anomalous results.” The adding-up test can be seen as a similar call to researchers to identify when and why these seemingly anomalous results arise and to expand our traditional theory and/or elicitation methods to include them.

3. Potential Difficulties in Implementing the Adding-Up Test

There are several potential difficulties that must be addressed in implementing an adding-up test. Haab et al. (2013) describe these issues and seem to suggest that the potential problems are so great that they outweigh the potential benefits of the test. We believe that these issues need to be considered on a case-by-case basis. In the

paragraphs below, we describe these potential difficulties and how they are addressed in our application.

Cognitive burden. The test requires that one part of the package of benefits be valued by respondents who are told that they already received another part. In many situations, this type of conditioning can be difficult for respondents to understand. In our application, we have been able to avoid this potential difficulty. One of the reasons we chose the Chapman et al. study is that its design is amenable to descriptions of incremental parts. As discussed below, the surveys for the incremental parts are the same from the respondents' perspective as the survey for the whole. No additional cognitive burden is imposed. In the original study for the base program (the whole), the years in which recovery will occur with and without the proposed intervention were stated to respondents. We simply changed these stated years for each of the incremental parts. In fact, this change in stated years was used in the original study for differentiating its scope and base versions, which gave us the idea that other increments could be defined similarly. In other applications, describing increments might be more difficult. But it can be useful to identify studies, like Chapman et al., in which the increments can be described without undue additional burden, and to apply adding-up tests in these applications.

Income effects. Ideally, respondents who are asked to evaluate the remaining part of the benefit bundle would have already paid for and received the first part of the bundle. The income effect is the recognition that if the respondent has already paid for the first part, then the available income on which he will state his WTP for the second part is reduced by the amount paid for the first part. Implementing such a payment is difficult and perhaps impossible in a survey setting. However, empirical methods can be applied

to address the issue. We apply these methods in our application, as did Bateman et al. (1997) in theirs. Note that if there were no income effects on WTP, then not conditioning on the payment does not affect the results of the analysis. In many situations, respondents' WTP for the goods in question are sufficiently small relative to their income such that income effects can reasonably be assumed to be negligible within that range (Diamond 1996). If potential income effects were a concern, the relation of respondents' incomes to their survey responses can be estimated. If income effects were found to exist, then the adding-up test can be implemented twice: once with the original responses and once with responses that are predicted at lower income levels to represent the WTP for the parts that are conditioned upon. In our application, we predict the votes at a lower income for each respondent. Since the estimated income effects are sufficiently small in our application, the predicted and actual votes are the same, such that the prediction under lower income did not change the results of the adding-up test.

Provision mechanism. Respondents might value a prospective good differently based on the way that a prior good is provided. For example, a prior good provided by nature can be viewed differently than the same good provided through human intervention, and this difference might affect the respondents' WTP for a prospective good.[viii] In our application, the prior and prospective goods are both provided by government programs (though different kinds of programs), and so there is less difference in the provision mechanism than between nature and human intervention. Also, in the original survey, the base program (the whole) was conditioned on government programs that provided prior benefits, with this conditioning described to respondents; the conditioning for the increments in our study takes the same form.

As Diamond (1996) originally pointed out, if respondents did indeed value a prospective good differently based on the provision method for a prior good, then their preferences would not be consistent with standard utility theory. In contrast, Hanemann (1994), e.g., argues that any factor may be a permissible element of consumers' utility. He does not, however, describe how normative allocation procedures can be derived in such an economic system.

Cost. The adding-up test is usually more expensive to apply than a scope test because it requires at least one more subsample. Fielding the survey is only one element of the overall cost of a project, and so a study with, e.g., three subsamples is not fifty percent more expensive than a study with two subsamples. In our application, the cost of fielding one additional subsample increased the overall cost by less than 5%. Given that the adding-up test potentially addresses the expert panel's concern about adequate response while the scope test does not, and that the burden of meeting the panel's concern "must rest" with the researcher, the extra cost seems justified, at least in some studies.

4. Review of Past Studies of Adding-up on Incremental Parts

We searched the natural resource valuation literature and could find only four studies whose designs permit an adding-up test on incremental parts: Samples and Hollyer (1990), Binger, Copple, and Hoffman (1995a and 1995b), Diamond et al. (1993), and Bateman et al. (1997). In the first two of these, the authors did not test for statistical significance or present results that allow readers to perform it. The first three studies are for public goods using CV, and the fourth is for private goods using an experimental setting for elicitation of WTP. We describe all four below.

Samples and Hollyer (1990) investigated whether the presence of substitutes or complements affected WTP values. In their design, respondents are first asked their WTP to save one type of marine mammal from a fatal disease and subsequently asked the additional amount they would pay to save a second type of mammal from the same disease, assuming that the first mammal is saved. A separate sample is asked their WTP to save both types of marine mammals simultaneously. They found that the sum of the WTPs for each mammal when asked incrementally greatly exceeded the WTP for the two mammals when asked about them together. Samples and Hollyer (1990) do not report the necessary statistical information to determine whether the difference is statistically significant.

Binger, Copple, and Hoffman (1995a and 1995b) also utilized a design that is amenable to an adding-up test. Their questionnaire first tells respondents about 57 different wilderness areas in four western states. In split samples, one group of respondents is first asked for their willingness to protect a specified wilderness area from timber harvests. Subsequently, that same group of respondents is asked for their WTP to protect the additional 56 wilderness areas, assuming that the first area is already protected. A separate sample of respondents is asked for their WTP to protect all 57 wilderness areas. They find that the sum of the average WTPs obtained from the first sample exceeds the average WTP from the second sample. However, like Samples and Hollyer (1990), these authors do not provide the necessary information that would reveal whether the difference is statistically significant.

Diamond et al. (1993) administered CV questionnaires to split samples of respondents that elicited their WTP to preserve specific wilderness areas, controlling for incremental parts by offering the various split samples different numbers of wilderness areas that are

already being developed. Their design allowed for two different adding-up tests, one with two parts and one with three parts. The results of the tests are mixed. The two-part test passes (the incremental parts add up to the total) while the three-part test fails (the incremental parts do not add up to the total).

In summary, of the three studies using CV on public goods, all three found that the sum of WTP for the incremental parts exceeded the WTP for the whole. Only one of the three studies (Diamond et al.) tested whether the difference was statistically significant, finding that the adding-up test failed in their three-part test and passed in their two-part test.

In addition to the three studies of public goods, there has been one study of adding-up of incremental parts with private goods. Bateman et al. (1997) used bidding in an experimental laboratory setting to measure respondents' WTP or WTA for vouchers for two components of a meal (the main course and dessert). Respondents were given an endowment of money and vouchers (one, two or none). To elicit WTP for a voucher, respondents who had not been given that voucher were told that they would need to state their WTP and then a random number would be drawn as the price of the voucher; if their stated WTP exceeded the randomly drawn price, then they would obtain the voucher at that price. WTP for both vouchers and WTA were elicited similarly. Four adding-up tests were applied based on WTP and WTA in each direction of conditioning. In all four comparisons they found that the sum of the WTP/WTA for the vouchers individually, when treated incrementally, exceeded the WTP/WTA for the two vouchers together. The difference was statistically significant for three of the comparisons (rejecting adding-up) and not significant for the fourth (one of the WTP comparisons).[ix]

As is the case for CV, the failures of adding-up found by Bateman et al. can be attributed to the elicitation method or because consumers' preferences do not adhere to the adding-up condition. Regarding the elicitation method, their experimental design might have introduced an effect that is similar to the "warm glow" that can arise in CV.[x] In particular, respondents may obtain some enjoyment from winning vouchers in each bid, independent of the value of the vouchers themselves.[xi] The small amount of money being bid, the fact that each respondent was given money by the experimenter to spend in bidding, and the use of random draws to determine whether the respondent wins, contribute to a game-like quality of the exercise.[xii] This warm glow of winning vouchers would cause the adding-up test to fail even if the true values of the vouchers themselves adhere to the adding-up condition. Alternatively, their results, as the authors say (p. 331), "may be a symptom of some fundamental property of individuals' preferences which conventional consumer theory does not allow for."

The amount by which the sum of the parts exceeds the whole is substantially smaller in Bateman et al. than in the studies, including ours, of CV for public goods. Bateman et al. find that the sum of the parts exceeded the whole by 5.3% to 16% in their experimental bidding for private goods, while we find that the sum of the parts in our CV study of a public good exceeds the whole by more than 200%. This comparison suggests that deviations from the adding-up condition – whether they arise from the elicitation method or from true preferences – are less severe with experimental bidding for private goods than with CV methods for public goods. More research is needed on adding-up, for both private and public goods and (if possible) with different elicitation methods, to assess the reasons and magnitudes of deviations from the adding-up condition.

5. The Original Study

The study of Chapman et al. (2009, hereafter "the Study") provides the basis for implementing the adding-up test. It was conducted by some of the most experienced researchers in the field and was funded at a sufficient level (over $2 million) to allow extensive design and revision of the various aspects of the study, including focus groups and pretesting of the instruments. It followed the procedures suggested by Carson and Groves (2007) that are intended to induce truthfulness; indeed, it is one of only three CV studies (that test for sensitivity to scope) to date to do so.[xiii] Its results served as the basis of expert testimony about damages in a court case, which is one of the most prominent purposes for which natural resource damages are calculated.

The goal of the Study was "to measure natural resource damages associated with excess phosphorus from poultry waste and other sources entering the Illinois River system [within Oklahoma] and Tenkiller Lake." The phosphorus creates excess algae that deplete the oxygen in the water, which is needed by aquatic species to survive. Respondents were informed that the state was taking measures to stop the spreading of poultry litter but that this action would not restore the lake and river[xiv] for a considerable period of time. Respondents were told that restoration could be hastened by putting alum (described as a naturally occurring mineral that is safe for humans) on the land and in the water, which binds to the phosphorus, rendering it harmless. In a referendum-type question, respondents were asked about their WTP for a program of alum treatment.

Two scenarios were specified and administered to two separate sets of respondents. For the "base" scenario, respondents were told that the ongoing actions of the state to reduce further pollution would restore the river to a natural state in 50 years and restore the lake in 60 years. With alum treatments in addition to these actions, the river would be

restored in 10 years instead of 50, which is 40 years earlier, and the lake would be restored in 20 years instead of 60, which is also 40 years earlier.

A “scope” scenario with reduced benefits was specified for the purposes of a standard scope test. Under the scope scenario, the impact of the state’s current actions and the alum program were both specified differently than in the base scenario. Respondents were told that the state’s current actions would restore the river in 10 years and the lake in 60 years. The alum treatment was for the lake, making its recovery “somewhat faster.” In particular, respondents were told that, with alum treatment of the lake, the lake would be restored in 50 years instead of 60 years, which is 10 years earlier. Note that the accelerated river restoration, which in the base scenario occurred as a result of the proposed alum treatments, occurs in the scope scenario as part of the state’s current actions.

Given the base and scope scenarios, the Study’s design represents three incremental parts:

A. Restoration of the river in 10 years instead of 50.
B. Restoration of the lake in 50 years instead of 60, given A.
C. Restoration of the lake in 20 years instead of 50, given A and B.

The base scenario is A, B, and C combined, and the scope scenario is B with its conditioning on A described to respondents.

6. Adaptation for Adding-Up Test

We expanded the number of increments from 3 to 4 for the following reason. Note that the Study’s scope scenario provides only 10 years of faster restoration starting 50 years

in the future. We were interested in whether respondents can differentiate distant times in their valuations. To address this question, we created another increment of lake restoration that provides only 10 years of faster restoration (like the scope scenario) but starts in 40 years instead of 50 years. Table 1 describes the resulting set of scenarios, and Figure 1 depicts them graphically.

For the "whole" and 2nd increment, we used the Study's survey instruments. For the 1st, 3rd, and 4th increments, we modified its instruments as little as possible to represent these situations.[xv] As discussed above, an important issue in adding-up tests is how to describe to respondents the conditioning on prior parts. We used the procedure that the Study utilized for the 2nd increment (its scope scenario). In particular, the conditioning for each increment is straightforward with this Study because respondents are already told that the state's current actions to prevent further pollution will restore the river and lake in some stated number of years for each. The numbers of years are changed for each version of the instrument to represent the conditioning. For the "whole" and the 1st increment, respondents are told 50 years for the river and 60 years for the lake. For the 2nd increment, the years are 10 and 60, respectively. For the 3rd increment, 10 and 50 years. And for the 4th increment, 10 and 40 years.

Note that the conditioning in each increment provides the same service as in the "whole" (accelerated river and/or lake restoration), but not through exactly the same form of provision as in the "whole" (the state's actions to prevent pollution rather than alum treatments). Both forms of provision are through actions by the state, but the prior increments are obtained through current government actions while the prospective ones are obtained through the new alum program.

We administered the questionnaires through the Internet, a procedure that is increasingly common in nonmarket valuation surveys (Berrens et al. 2004; Banzhaf et al. 2006; Fleming and Bowden 2009; Windle and Rolfe 2011). In addition to the lower cost relative to in-person interviews, Internet surveys have the advantage of seamless incorporation of diagrams, photos, and other visual aids. Our practice differs from the Study, which conducted in-person interviews. The difference largely reflects a difference in purpose. The Study was estimating damages for litigation purposes, for which the sample needs to be representative of the target population. Our purpose is to assess whether CV responses are adequately sensitive to differences in scope, and our findings are relevant at least to our experimental samples.

We took several steps to adapt the Study's in-person questionnaire to an Internet survey. First, we conducted 105 cognitive, in-person interviews, using several versions of the questionnaire, to better appreciate how people answered the questions, which informed our structuring of the on-line versions. Secondly, we pre-tested two on-line versions of the questionnaire with 79 respondents, all of whom were able to complete the survey without the aid of an interviewer. Third, we added an opportunity for the on-line respondents to provide open-ended comments at the end of the questionnaire, and nearly all of these open-ended responses indicated that the respondents considered the questionnaire to be understandable and enjoyable.

To implement the adding-up test, we fielded five versions of the questionnaire, which are described in Table 1. We randomly assigned to each respondent one of the six bids ($10, $45, $80, $125, $205, and $405) used in the Study, as well as one of the five versions.

The surveys were fielded between November 2011 and March 2012. For our primary analysis, we excluded some responses. First, we excluded respondents who spent less than 15 minutes or more than 120 minutes completing the survey. In the first case, we did not believe that respondents could carefully consider the full content of the questionnaire in such a short amount of time. When respondents took more than 120 minutes to complete the questionnaire, we believed that they likely walked away from their computer during the course of the survey, such that we could not know whether they had actually spent at least 15 minutes on the task. We also eliminated the 14 respondents who did not answer the open-ended question about why they voted for or against the proposed program to accelerate restoration because we were concerned that including respondents who gave no reasons could bias the results against a finding of adequate, or reasonable, response to scope. After these eliminations, the primary analysis contained 980 responses across the five versions.

We compared the subsamples to determine whether there were significant differences among them. The demographic characteristics of each subsample are given in Table 2. On visual inspection, the subsamples seem to be similar, as would be expected from the fact that respondents were selected randomly for the subsamples. We performed one-way ANOVA tests of equality of the demographic means across subsamples. In all cases, the hypothesis of no difference could not be rejected at usual confidence levels.

7. Results

We first report on the traditional scope tests for each of the four increments (separately) relative to the whole. We use the same non-parametric approach used by the Study. Specifically, we compare the percentage of respondents who voted for the program at

each bid and use a Wald test to test jointly whether the differences are statistically significant. Table 3 shows the details. The hypothesis of no difference is rejected twice (for the 2nd and 3rd increments, which pass the scope test)[xvi] and accepted twice (for the 1st and 4th increments, which fail the scope test).

To estimate WTP associated with each of the versions, we used the ABERS non-parametric estimator (Ayer et al. 1955), the same as the Study. Table 4 shows the summary statistics for each of the versions.[xvii] Given the interval nature of the data, we used bootstrapping techniques (Efron 1982; Davison and Hinkley 1997) to determine whether these WTP values are statistically different from each other. The WTP differences are consistent with the test of proportions in Table 3 above. Specifically, WTP for the whole is statistically different from the WTP for the 2nd and 3rd increments and is not statistically different from the WTP for the 1st and 4th increments.

The adding-up test is based on the mean WTP values displayed in Table 4. The sum of the four increments totals $609 (=187+97+144+181), which is about three times as large as the value of the whole ($200). We applied the bootstrap to simulate the sampling distribution of the difference between the mean WTP for the whole and the sum of the mean WTP from the four increments. The 99-percent confidence interval does not contain zero, such that the hypothesis of equality is rejected: the responses fail the adding-up test.

We conducted several types of sensitivity analyses. To investigate whether our results would change if sample sizes were larger, we re-fielded the 2nd increment version, which was the Study's scope version, with a larger sample size: nearly 500 respondents after exclusions. The mean WTP for this re-fielded version is $103, which is not statistically

different from the $97 in Table 4. We also included respondents who took 10 to 14 minutes to complete the survey, which increased the sample size across the five versions from 980 to 1,106. With these higher sample sizes, our results do not change. Finally, we also applied post-stratification weights to our respondents' answers to reflect the population in terms of gender, age, and education. This weighting does not change the results of the adding-up test. With the weighted data, the ratio of the sum of the parts to the whole is still larger than 3 to 1.

As discussed above, the 2nd and 3rd increments both provide 10 years of faster lake restoration, but starting at different times in the future. If consumers discount appropriately, the 3rd increment should be valued more than the 2nd increment, since the benefits in the 3rd increment start sooner than those in the 2nd increment. The estimates in Table 4 conform to this expectation. These values are not statistically different from each other at the 95 percent level but are different at the 90 percent level.

8. Income Effects

As discussed above, in an ideal adding-up test, the incremental specification of benefits would reduce respondents' income by their WTP for prior parts before they evaluate a prospective part. If there were no income effects in the relevant range, then not reducing respondents' incomes does not affect their valuations. We tested for the existence of income effects. In particular, we estimated binary logit models of whether the respondent answered "yes" or "no" to the referendum question (i.e. voted for or against the program at the specified cost.) We included the cost that the respondent faced, as well as income and other demographics. The results are given in Table 5 for the entire sample. Income enters with a t-statistic of 0.95, such that the hypothesis of no income effects cannot be rejected. The point estimate of the impact of income on

response is exceedingly small. We also estimated the model for each subsample separately. In all models (not shown), the income coefficient was insignificant. The point estimate was positive in four of the subsamples and negative in one, and very small in magnitude in all subsamples.

We used the estimated model in Table 5 to simulate the impact of a decrease in income for the respondents who faced an increment that conditioned on a prior increment, i.e., who faced the 2nd, 3rd, and 4th increments. (Respondents who faced the whole and the 1st increment had no prior benefits upon which to condition.) The simulation was performed as follows.

Let $U_n(y_n)$ be person *n*'s utility from the benefits that were described to the person, net of the cost that was specified, given an income of y_n.[xviii] As usual for derivation of choice models, utility is decomposed into a part observed by the researcher and an unobserved part: $U_n(y_n) = V_n(y_n) + \varepsilon_n$. Assuming that ε_n is distributed logistic, the probability that the person votes "yes" is

$$P_n(y_n) = Prob(U_n(y_n) > 0) = Prob(\varepsilon_n > -V_n(y_n)) = \frac{1}{1+e^{-V_n(y_n)}}.$$

This probability was used for the model in Table 5, which gives the estimate of $V_n(y_n)$.

Consider now the person's choice if the original income is lower by deduction *d*. Utility is $U_n(y_n - d) = V_n(y_n - d) + \varepsilon_n$ and the probability of voting "yes" is $P_n(y_n - d) = \frac{1}{1+e^{-V_n(y_n-d)}}$. This is the unconditional probability; however, we observe whether respondents voted "yes" or "no" at their original income, and this information can used to provide a better estimate of the probability of voting "yes" at the lower income. For respondents who voted "no" at their original income, the conditional

probability of voting “yes” at a lower income is zero (assuming income effects are nonnegative.) For respondents who voted “yes” at their original income, the conditional probability of voting “yes” at a lower income is

$$Prob(U_n(y_n - d) > 0 | U_n(y_n) > 0)$$

$$= \frac{Prob(U_n(y_n - d) > 0) * Prob(U_n(y_n) > 0 | U_n(y_n - d) > 0)}{Prob(U_n(y_n) > 0)}$$

$$= \frac{Prob(U_n(y_n - d) > 0)}{Prob(U_n(y_n) > 0)} = \frac{1 + e^{-V_n(y_n)}}{1 + e^{-V_n(y_n - d)}}$$

We calculated the conditional probabilities of voting “yes” if *d* were deducted from the income for all respondents who voted “yes” at their original income. We then simulated each of these respondents’ votes by taking a draw from a uniform distribution and changing the “yes” vote to “no” if the draw for that respondent exceeded the conditional probability. We set *d=$1000*, which is far greater than the largest cost that was presented to anyone.

Among respondents who voted “yes” at their original income, the conditional probability of voting “yes” at the lower income is, on average, 99.89 percent. With such a high probability, no respondents were simulated to change their vote from “yes” to “no” at the lower income. This result, of course, is due to the fact that the estimated income effect is so small. If there had been a difference between the simulated and original votes, then the adding-up test could be applied to the simulated votes and the results compared to those, described above, for the original votes.

9. Discussion

The scope test has been applied as a means to ascertain whether CV results reflect economic preferences. However, passing the scope test does not imply that the magnitude of the estimated response is adequate, and scope test failures can be explained by certain conditions that are consistent with economic theory. As an additional step to resolve these uncertainties, and to address the NOAA expert panel's concern about adequate response to scope changes, we recommend the adding-up test of Diamond et al. (1993).

Building on a CV study by Chapman et al. (2009) that already contained incremental parts, we expanded the study to contain a full set of incremental parts and then applied an adding-up test. We found that the adding-up condition does not hold for the CV results: the sum of the estimated WTPs for the incremental parts greatly exceeds the estimated WTP for the whole. Our results mirror the conclusions of Diamond et al. (1993) using CV on public goods and Bateman et al. (1997) using a laboratory setting on private goods.

As discussed above, failure of the adding-up test in our study can indicate that the CV procedure is not obtaining truthful answers from respondents, and/or that consumers' preferences are not consistent with standard utility theory. In regard to truthfulness, unlike previous studies of adding-up for public goods, we used CV scenarios that are consequential, as described by Carson and Groves (2007), and therefore designed to induce truthful answers from respondents. So either the Carson and Groves procedures do not actually induce truthfulness, or respondents' truthful answers are not consistent with the adding-up condition.

Behavioral theories may be useful in understanding the sources and patterns of responses and might provide a behavioral explanation for failures of the adding-up test. Bateman et al. (2004), Powe and Bateman (2004), and Heberlein et al. (2005) provide explanations for scope test failures that might also be applicable to the adding-up test. As well as developing the explanations, steps are needed to derive an expanded theory of welfare that incorporates these explanations, or elicitation methods that avoid the behaviors.

Bateman (2011) suggests tests that complement the adding-up test and could be explored. Diamond (1996) proposed methods based on properties of the second derivatives of utility, which, to our knowledge, have not been implemented in empirical work. We endorse more research along these lines to develop and apply other tests of the consistency of responses with standard utility theory and, insofar as inconsistencies are found, to develop methods that account for them and theories that explain them.

References

Alvarez-Farizo, Begona, Nick Hanley, Robert E. Wright, and Douglas Macmillan. 1999. "Estimating the Benefits of Agri-environmental Policy: Econometric Issues in Open-ended Contingent Valuation Studies." *Journal of Environmental Planning and Management* 42(1): 23 – 43.

Arrow, K., R. Solow, P.R. Portney, E.E. Leamer, R. Radner, and H. Schuman. 1993. "Report of the NOAA Panel on Contingent Valuation." 58 *Fed. Reg*. 4601 *et. seq*. Jan. 15.

Arrow, Kenneth, Edward E. Leamer, Howard Schuman, and Robert Solow. 1994. "Comments of Proposed NOAA Scope Test." Appendix D of *Comments of Proposed NOAA/DOI Regulations on Natural Resource Damage Assessment*, U.S. Environmental Protection Agency.

Ayer, Miriam, H. D. Brunk, G.M. Ewing, W.T. Reid, and Edward Silverman. 1955. "An Empirical Distribution Function for Sampling with Incomplete Information." *The Annals of Mathematical Statistics* 26(4): 641–647.

Banzhaf, H. Spencer, Dallas Burtraw, David Evans, and Alan Krupnick. 2006. "Valuation of Natural Resource Improvements in the Adirondacks." *Land Economics* 82(3): 445–464.

Bateman, Ian J. 2011. "Valid Value Estimates and Value Estimate Validation: Better Methods and Better Testing for Stated Preference Research." In *The International Handbook on Non-Market Environmental Valuation*, ed. Jeff Bennett. Cheltenham UK: Edward Elgar Publishing.

Bateman, Ian J., M.P. Cameron, and A. Tsoumas. 2008. "Investigating the Characteristics of Stated Preferences for Reducing the Impacts of Air Pollution: A Contingent Valuation Experiment." In *Environmental Economics, Experimental Methods*, eds. Todd L. Cherry, Stephan Kroll, and Jason F. Shogren. London: Routledge.

Bateman, Ian J., Matthew Cole, Philip Cooper, Stavros Georgiou, David Hadley, and Gregory L. Poe. 2004. "On Visible Choice Sets and Scope Sensitivity." *Journal of Environmental Economics and Management* 47: 71–93.

Bateman, Ian J, A. Munro, B. Rhodes, C. Starmer, and R. Sugden. 1997. "Does Part Whole Bias Exist? An Experimental Investigation." *Economic Journal* 107(441): 322-332.

Berrens, Robert P., Alok K. Bohara, Hank C. Jenkins-Smith, Carol L. Silva, and David L. Weimer. 2004. "Information and Effort in Contingent Valuation Surveys: Application to Global Climate Change Using National Internet Samples." *Journal of Environmental Economics and Management* 47: 331–363.

Binger, Brian, Robert Copple, and Elizabeth Hoffman. 1995a. "Contingent Valuation Methodology in the Natural Resource Damage Regulatory Process: Choice Theory and the Embedding Phenomenon." *Natural Resources Journal* 35(3): 443–459.

Binger, Brian R., Robert F. Copple, and Elizabeth Hoffman. 1995b. "The Use of Contingent Valuation Methodology in Natural Resource Damage Assessments: Legal Fact and Economic Fiction." *Northwestern University Law Review* 89(3): 1029–1053.

Boyle, Kevin J., William H. Desvousges, F. Reed Johnson, Richard W. Dunford, and Sara P. Hudson. 1994. "An Investigation of Part-Whole Biases in Contingent Valuation Studies." *Journal of Environmental Economics and Management* 27: 64-83.

Carson, Richard T. 1997. "Contingent Valuation and Tests of Insensitivity to Scope." In *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods,* eds. R.J. Kopp, W. Pommerhene, and N. Schwartz. Boston: Kluwer.

Carson, Richard T. and Theodore Groves. 2007. "Incentive and Informational Properties of Preference Questions." *Environmental and Resource Economics* 37:181–210.

Carson, Richard and W. Michael Hanemann. 2005. "Contingent Valuation." In *Handbook of Environmental Economics, Vol. 2*, eds. K.-G. Mäler and J. R. Vincent, 822–935. Amsterdam: Elsevier.

Carson, Richard T., W. Michael Hanemann, Raymond J. Kopp, Jon A. Krosnick, Robert C. Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith. 1994. *Prospective Interim Lost Use Value Due to DDT and PCB Contamination in the Southern California Bight.* Report prepared for the National Oceanic and Atmospheric Administration. La Jolla, CA: Natural Resource Damage Assessment, Inc. September 30.

Chapman, David J., Richard C. Bishop, W. Michael Hanemann, Barbara J. Kanninen, Jon A. Krosnick, Edward R. Morey, and Roger Tourangeau. 2009. "Natural Resource Damages Associated with Aesthetic and Ecosystem Injuries to Oklahoma's Illinois River System and Tenkiller Lake." Expert Report for State of Oklahoma, Volume I, available at https://pcl.uscourts.gov/search (*Oklahoma v. Tyson Foods Inc.*, No. 4:05-cv-329 (N.D. Okla. Feb. 13, 2009), Docket No. 1853-4, exhibit D). The public-use copy from this site is also available at http://elsa.berkeley.edu/~train/chapman.pdf

Christie, Michael. 2001. "A Comparison of Alternative Contingent Valuation Elicitation Treatments for the Evaluation of Complex Environmental Policy." *Journal of Environmental Management* 62(3): 255–269.

Davison, A.C. and D.V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Desvousges, William, Kristy Mathews, and Kenneth Train. 2012. "Adequate Response to Scope in Contingent Valuation." *Ecological Economics* 84: 121-128.

de Zoysa, A. Dimanta N. 1995. "A Benefit Valuation of Programs to Enhance Groundwater Quality, Surface Water Quality, and Wetland Habitat in Northwest Ohio." PhD. Dissertation. The Ohio State University.

Diamond, Peter A. 1996. "Testing the Internal Consistency of Contingent Valuation Surveys." *Journal of Environmental Economics and Management* 30: 337–347.

Diamond, Peter A., Jerry A. Hausman, Gregory K. Leonard, and Mike A. Denning. 1993. "Does Contingent Valuation Measure Preferences? Experimental Evidence." In *Contingent Valuation, A Critical Assessment*, ed. J.A. Hausman, 41–89. Amsterdam: Elsevier.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

Fleming, Christopher M. and Mark Bowden. 2009. "Web-based Surveys as an Alternative to Traditional Mail Methods." *Journal of Environmental Management* 90: 284–292.

Haab, Timothy C., Matthew G. Interis, Daniel Petrolia, and John C. Whitehead. 2013. "From Hopeless to Curious? Thoughts on Hausman's `Dubious to Hopeless' Critique of Contingent Valuation." *Applied Economic Perspectives and Policy* 35(4):593-612.

Hanemann, W. Michael. 1994. "Valuing the Environment Through Contingent Valuation." *The Journal of Economic Perspectives* 8(4):19 - 43.

Heberlein, Thomas A., Matthew A. Wilson, Richard C. Bishop, and Nora Cate Schaeffer. 2005. "Rethinking the Scope Test as a Criterion for Validity in Contingent Valuation." *Journal of Environmental Economics and Management* 50: 1–22.

Hoevengel, Ruud. 1996. "The Validity of the Contingent Valuation Method: Perfect and Regular Embedding." *Environmental and Resource Economics* 7: 57–78.

Loomis, John and Armando Gonzalez-Caban. 1998. "A Willingness-to-Pay Function for Protecting Acres of Spotted Owl Habitat from Fire." *Ecological Economics* 25: 315–322.

Macmillan, D.C. and E.I. Duff. 1998. "Estimating the Non-Market Costs and Benefits of Native Woodland Restoration Using the Contingent Valuation Method." *Forestry* 71 (3): 247–259.

Nunes, Paulo A.L.D. and Erik Schokkaert. 2003. "Identifying the Warm Glow Effect in Contingent Valuation." *Journal of Environmental Economics and Management* 45: 231–245.

Powe, Neil A. and Ian J. Bateman. 2004. "Investigating Insensitivity to Scope: A Split-Sample Test of Perceived Scheme Realism." *Land Economics* 80(2): 258–271.

Riddel, Mary and John Loomis. 1998. "Joint Estimation of Multiple CVM Scenarios under a Double Bounded Questioning Format." *Environmental and Resource Economics* 12: 77–98.

Rollins, Kimberly and Audrey Lyke. 1998. "The Case for Diminishing Marginal Existence Values." *Journal of Environmental Economics and Management* 36: 324–366.

Samples, K.C and J.R. Hollyer. 1990. "Contingent Valuation of Wildlife Resources in the Presence of Substitutes and Complements." In *Economic Valuation of Natural Resources: Issues, Theory, and Applications*, eds. R.L. Johnson and G.V. Johnson, 177–192. Boulder, CO: Westview Press.

Stevens, T.H., S. Benin, and J.S. Larson. 1995. "Public Attitudes and Values for Wetland Conservation in New England." *Wetlands* 15(3): 226–23.

Streever, W.J., M. Callaghan-Perry, A. Searles, T. Stevens, and P. Svoboda. 1998. "Public Attitudes and Values for Wetland Conservation in New South Wales, Australia." *Journal of Environmental Management* 54(1): 1–14.

Veisten, Knut, Hans Fredrik Hoen, and Jon Strand. 2004. "Sequencing and the Adding Up Property in Contingent Valuation of Endangered Species: Are Contingent

Non-use Values Economic Values?" *Environmental and Resource Economics* 29: 419-423.

White, Piran C. L., Keith W. Gregory, Patrick J. Lindley and Glenn Richards. 1997. "Economic Values of Threatened Mammals in Britain: A Case Study of the Otter *Lutra Lutra* and the Water Vole *Arvicola Terrestris." Biological Conservation* 82: 345–354.

Whitehead, John C., Timothy C. Haab, and Ju-Chin Huang. 1998. "Part-Whole Bias in Contingent Valuation: Will Scope Effects Be Detected with Inexpensive Survey Methods?" *Southern Economic Journal*, 65(1):160-168.

Windle, Jill and John Rolfe 2011. "Comparing Responses from Internet and Paper-Based Collection Methods in more Complex Stated Preference Environmental Valuation Surveys." *Economic Analysis and Policy* 41(1): 83–97.

Wu, Pei-Ing. 1993. "Substitution and Complementarity in Commodity Space: Benefit Evaluation of Multidimensional Environmental Policy." *Academia Economic Papers* 21(1): 151–182.

TABLES

Table 1: Questionnaire Versions

Version	Description
Whole: River + Lake	WTP for accelerating the restoration of the lake from 60 years to 20 years (40 years sooner) and accelerating the restoration of the river from 50 years to 10 years (40 years sooner), given that the state's current actions will induce the river to be restored in 50 years and the lake to be restored in 60 years
1st Increment: River	WTP for accelerating the restoration of the river from 50 years to 10 years (40 years sooner), given that the state's current actions will induce the river to be restored in 50 years and the lake to be restored in 60 years
2nd Increment: Lake 10 years	WTP for accelerating the restoration of the lake from 60 years to 50 years (10 years sooner), given that the state's current actions will induce the river to be restored in 10 years and the lake to be restored in 60 years
3rd Increment: Lake 10 more years	WTP for accelerating the restoration of the lake from 50 years to 40 years (10 years sooner), given that the state's current actions will induce the river to be restored in 10 years and the lake to be restored in 50 years
4th Increment: Lake 20 more years	WTP for accelerating the restoration of the lake from 40 years to 20 years (20 years sooner), given that the state's current actions will induce the river to be restored in 10 years and the lake to be restored in 40 years

Table 2: Demographic Variables by Subsample

	Subsample				
Demographics	Whole	1st Increment	2nd Increment	3rd Increment	4th Increment
Percent Male	33%	27%	33%	34%	29%
Percent College	28%	28%	30%	27%	34%

	Subsample				
Demographics	Whole	1st Increment	2nd Increment	3rd Increment	4th Increment
Graduate					
Percent Strong Environmentalist	14%	15%	8%	12%	13%
Average Age	46	49	47	48	48
Average Income	$42,900	$44,300	$43,700	$40,400	$43,800
Sample Size	172	293	159	174	182

Table 3: Percent of Respondents Voting for the Alum Treatments

Bid	**Whole:** River+Lake	**1st Increment:** River	**2nd Increment:** Lake 10 years	**3rd Increment:** Lake 10 more years	**4th Increment:** Lake 20 more years
10	68.0%	74.5%	50.0%	82.8%	72.7%
45	60.6%	58.3%	37.5%	48.1%	44.0%
80	69.2%	64.6%	29.2%	32.3%	64.9%
125	50.0%	57.4%	42.9%	23.1%	62.5%
205	44.8%	38.9%	24.0%	40.7%	35.7%
405	45.2%	40.0%	15.4%	35.3%	40.7%
Wald Test F Statistic		0.23	3.98	2.71	0.56
P-Value		0.9674	0.0007	0.0139	0.7634

Table 4: WTP Estimates

Version	**Mean WTP**	**Standard Error**	**95% Confidence Interval**
Whole: River + Lake	$200	17.71	$165 - $235
1st Increment: River	$187	12.31	$163 - $211
2nd Increment: Lake 10 years	$97	13.73	$ 70 - $124
3rd Increment: Lake 10 more years	$144	15.34	$114 - $174
4th Increment: Lake 20 more years	$181	18.69	$144 - $218

Table 5: Logit Model of Yes/No Vote

	Estimated coefficient	**Standard error**
Cost, in dollars	-0.0031	0.0005
Income, in thousands of dollars	0.0022	0.0023
Age	-0.0025	0.0046
Male	-0.1039	0.1506
College graduate	-0.3278	0.1636
Concerned about environment	1.0609	0.2205

	Estimated coefficient	Standard error
Whole	0.6744	0.2870
1st Increment	0.6680	0.2763
2nd Increment	-0.3791	0.2969
3rd Increment	0.1582	0.2908
4th Increment	0.6654	0.2875
Log-likelihood	-608.502	
Sample size	950	

FIGURES

Figure 1: Incremental Parts of Accelerated Restoration

[i] Boyle et al. (1994) were among the first to provide empirical evidence about the relevance of scope and its implications for the reliability of CV estimates, especially for non-use, or passive use, values.

[ii] "We believe that there is a very sharp contrast between the basic character of the proposed scope test and the sense of the NOAA panel. Because of this difference, we do not think that this test is a proper response to the Panel report.…The report of the NOAA panel calls for survey results that are 'adequately' responsive to the scope of the environment insult. The proposed scope test is built to assure that there is a statistically detectable sensitivity to scope. This is, in our opinion, an improper interpretation of the word 'adequately.' Had the panel thought that something as straightforward as statistical measurability were the proper way to define sensitivity, then we would (or should) have opted for language to that effect." (underlining in the original.)

[iii] One study, Bateman et al. (1997), applied the adding-up test to private goods with a bidding-based elicitation procedure in a laboratory setting. We discuss this study and its implications in Section 4.

[iv] The studies that examine adding-up on non-incremental parts include Alvarez-Farizo et al. (1999), Bateman, Cameron, and Tsoumas (2008), Christie (2001), de Zoysa (1995), Macmillan and Duff (1998), Nunes and Schokkaert (2003), Powe and Bateman (2004), Stevens, Benin, and Larson (1995), Veisten, Hoen and Strand (2004), White et al. (1997), and Wu (1993). The test is failed in all of these studies except Nunes and Schokkaert, who pass their adding-up test when they use a factor analysis to account for warm glow, and not otherwise. Other studies have designs that support an adding-up test on non-incremental parts, but the authors do not report the results (Hoevengal 1996; Loomis and Gonzalez-Caban 1998; Riddel and Loomis 1998; Rollins and Lyke 1998; Streever et al. 1998).

[v] De Zoysa's survey asked a referendum-style question, which is consistent with Carson and Groves' recommendations, but followed-up with an open-ended question asking respondents to state their maximum WTP, which violates Carson and Groves' concepts of consequentiality. If respondents did not anticipate that the follow-up was going to be asked when answering the referendum question (or did not read ahead before answering the referendum question in the mail survey), then the answers to the referendum question can be considered to be consistent with Carson and Groves' recommendations.

[vi] The adding-up condition does not contradict the fact that goods are often priced with bundled discounts, under which buying each good individually costs more in total than buying the goods as a bundle. The adding-up condition describes the amount that consumers are willing to pay, while the pricing mechanism describes the amount that consumers are required to pay. In fact, marketers exploit consumers' adding-up condition when offering bundled prices. E.g., suppose a consumer is willing to pay $7 for one unit of a good, and $5 for a second unit once the first unit is obtained. By the adding-up condition, the consumer is willing to pay $12 for two units. With non-bundled pricing, the seller can price at $7, sell 1 unit to this customer, and make $7 in revenue; or

price at $5, sell 2 units, and make revenue of $10. However, by offering a bundled price of two units for $12, the seller sells two units and obtains revenues of $12. If the consumer's WTP for the two units incrementally exceeded the amount the consumer is willing to pay for both units together (in violation of the adding-up condition), then the seller could not make as much, or any, extra revenue though bundling.

[vii] Using their notation, Whitehead, Haab and Huang define $\Delta WTP = WTP_{1,2} - WTP_2$ and show that $\Delta WTP = e(p_1, p_2, q_1, q_2^*, u) - e(p_1, p_2, q_1^*, q_2^*, u)$ and then $\Delta WTP \geq 0$ under the standard assumptions of utility theory. By the definition of WTP, this second equation shows that ΔWTP is the WTP for 1 given 2, which can be denoted $WTP_{1|2}$. Their first equation then becomes $WTP_{1|2} = WTP_{1,2} - WTP_2$, which is the adding-up condition. No new assumptions have been introduced.

[viii] It is not clear what the direction of effect would be: does valuing the prior good differently because of its provision method raise or lower the respondent's WTP for the prospective good? The different valuation of the prior good would need to raise the respondents' WTP of the prospective good in order to induce false failures of the adding-up test. The opposite would cause false acceptances of the adding-up test.

[ix] The authors recently corrected the t-statistics in their Table 3 (personal communication.) The corrected t-statistics, in order of the rows in Table 3, are 2.55, 0.96, 2.98, 2.23.

[x] "Warm glow" refers to the idea that respondents obtain satisfaction from expressing support for an environmental improvement, independent of their value of the improvement itself.

[xi] Where winning vouchers means getting or keeping the vouchers in each potential trade.

[xii] As the authors describe: "…each subject faced a screen, rather like a roulette wheel, around which were located a range of prices at which the trade might conceivably be carried out….A 'ball' then circled around the wheel and alighted at one sum at random." (p. 326)

[xiii] The others are Carson et al. (1994) and (possibly) de Zoysa (1995). Desvousges, Mathews, and Train (2012) identify several other papers that nearly adhere to the Carson and Groves procedures.

[xiv] For linguistic convenience, we refer to "Illinois River system within Oklahoma" as "the river."

[xv] The Study provided considerable background information to respondents to allow them to place the alum program in context. We provided the same background information to respondents of the increment versions. For interested readers, the survey used in the Study is described in the citation for Chapman et al. (2009) in the References section below. The instruments that we used are available from the authors on request.

[xvi] The Study's scope test is equivalent to our 2nd increment relative to the whole, which was passed, the same as we find.

[xvii] For the whole, the Study's mean is $184. For the 2nd increment, the Study's mean is $138. Our and the Study's confidence intervals overlap. The similarity of results suggests that our application of the survey in Internet form, and the passage of time since the original Study, did not materially affect the responses. It does not suggest that either set of responses is reliable as a measure of WTP, just that similar surveys induce similar responses.

[xviii] The utility function might take the form $U_n(y_n) = \alpha(WTP_n(y_n) - c_n)$ where c_n is the program costs that the person faced and WTP is random from the researcher's perspective.