Temporal Reliability of Estimates from Contingent Valuation

Richard T. Carson, W. Michael Hanemann,
Raymond J. Kopp, Jon A. Krosnick,
Robert C. Mitchell, Stanley Presser, Paul A. Ruud,
and V. Kerry Smith

with Michael Conaway and Kerry Martin

Discussion Paper 95-37

August 1995

1616 P Street, NW
Washington, DC 20036
Telephone 202-328-5000
Fax 202-939-3460

Temporal Reliability of Estimates from Contingent Valuation

Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Jon A. Krosnick,
Robert C. Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith

with

Michael Conaway and Kerry Martin

Abstract

In 1992 the National Oceanic and Atmospheric Administration (NOAA) convened a panel of prominent social scientists to assess the reliability of natural resource damage estimates derived from contingent valuation (CV). The product of the panel's deliberations was a report that laid out a set of recommended guidelines for CV survey design, administration, and data analysis. This paper focuses on one of these guidelines -- the Panel's call for the "temporal averaging" of willingness-to-pay (WTP) responses obtained from CV surveys as one method for increasing their reliability. The panel suggested: "Time dependent measurement noise should be reduced by averaging across independently drawn samples taken at different points in time. A clear and substantial time trend in the responses would cast doubt on the 'reliability' of the finding."

The purpose of this paper is to examine the temporal reliability of CV estimates. Our findings, using a CV instrument designed to measure willingness-to-pay for a program to protect Prince William Sound, Alaska from future oil spills, like the Exxon Valdez spill, exhibited no significant sensitivity to the timing of the interviews. For two samples involving independent interviews taken over two years apart, the distribution of respondents' choices "for" and "against" the protection program did not differ.

Key Words: contingent valuation, natural resource damages, passive use, Exxon Valdez, reliability

JEL Classification Nos.: D60, D61, K32, Q28

Table of Contents

List of Tables

Temporal Reliability of Estimates from Contingent Valuation

Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Jon A. Krosnick,
Robert C. Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith

with

Michael Conaway and Kerry Martin*

1. INTRODUCTION

Over the past two decades the use of contingent valuation (CV) in policy analysis and academic research has grown rapidly. According to one estimate there are now almost two thousand papers and studies in the literature dealing with CV [see Carson, Wright, Carson, Alberini and Flores, 1995]. Special attention has focused on using CV to estimate passive use.[1] For example, the U.S. Environmental Protection Agency's (EPA) recent cost-benefit study in support of a proposed environmental regulation under a hazardous substance law known as the Resource Conservation and Recovery Act (RCRA), explicitly focused on measuring economic values for cleaning up groundwater pollution with particular attention

* The authors are, respectively: Associate Professor of Economics, University of California (San Diego); Associate Professor of Agricultural and Natural Resource Economics, University of California (Berkeley); Senior Fellow, Resources for the Future; Associate Professor of Psychology and Political Science, Ohio State University; Professor of Geography, Clark University; Professor of Sociology, University of Maryland (College Park); Professor of Economics, University of California (Berkeley); and Arts and Sciences Professor, Duke University and University Fellow, Resources for the Future. Conaway and Martin are members of Natural Resource Damage Assessment, Inc. and made extensive contributions throughout the effort. The work described in this paper was funded by the Damage Assessment Office of the National Oceanic and Atmospheric Administration as part of a natural resource damage assessment under contract number 50-DGNC-1-00007. Additional support to aid in the preparation of this paper was provided to Kopp by the Alfred P. Sloan Foundation through its support of the Welfare Economics Program at Resources for the Future and to Smith by the UNC Sea Grant Program under Grant No. R/MRD-25. Thanks to Richard Bishop, Trudy Cameron, Nicholas Flores, Carol Jones, Norman Meade, Pierre Du Vair and Alan Randall for comments on aspects of this work. All opinions expressed in this paper are those of the authors and should not be attributed to the National Oceanic and Atmospheric Administration, the Alfred P. Sloan Foundation, or any persons or organizations acknowledged above.

[1] The term passive use was first used in the ruling by the United States Court of Appeals for the District of Columbia in *Ohio v. Department of the Interior,* 880 F.2d 432 (D.C. Cir. 1989). The value derived from passive use has been referred to as nonuse value, existence value, bequest value and option value.

given to measuring the nonuse (or passive use) values associated with groundwater protection [see McClelland *et al*., 1992]. Further, in an important judicial opinion (*Ohio v. Department of the Interior*), a three judge Court of Appeals ruling stated that lost passive use values should be included in damage awards resulting from injuries to natural resources due to releases of hazardous substances.[2] Under the *Ohio* decision, it is unnecessary for an individual to be a direct user of a natural resource, say a recreationist, to hold an economic value for the resource in question (or for some aspect of the services that may be provided by the resource).[3] The *Ohio* Court also emphasized the importance of the "reliability" of methods used to estimate natural resource damages.[4] Because contingent valuation is currently the only technique available to measure economic values that encompasses both use and passive use, much of the current CV research has been directed at evaluating its reliability.

---

[2] The opinion in *Ohio v. Department of the Interior* stated,

> On remand, DOI should consider a rule that would permit trustees to derive use values for natural resources by summing up all reliably calculated use values, however measured, so long as the trustee does not double count. [p. 87]

The opinion made clear that its definition of use values included use and passive use or nonuse values.

[3] We adopt the term "economic value" rather than simply "value" to distinguish our meaning from other uses of the word value. Economic value is defined in terms of individual choices. When someone chooses to give up x in order to obtain y, we can say that the economic value of y (termed the object of choice) is at least x.

[4] In the debate over the appropriate uses of CV, the word "reliability" is frequently used. It is not apparent, however, that this word has the same meaning to all the participants in the debate. As noted in Kopp and Pease [1995], a recent U.S. Supreme Court decision concerning the admissibility of scientific evidence (*Daubert v. Merrell Dow Pharmaceuticals*, 113 S.Ct. 2786, 2795, n9 (1993)), noted that while scientists "typically distinguish between 'validity' (does the principle support what it purports to show?) and 'reliability' (does application of the principle produce consistent results?)," the Court emphasized its "reference here is to evidentiary reliability -- that is, trustworthiness." As used by the *Ohio* Court and in the NOAA Panel report, the reliability of a measure is the degree to which it measures the theoretical construct under investigation. However, in the empirical social sciences, this preceding definition pertains to *validity*, whereas reliability is defined as the extent to which the variance of the measure is not due to random sources and systematic sources of error. In this paper we are using the term reliability in this latter sense.

To assess the reliability of natural resource damage estimates derived from CV, the National Oceanic and Atmospheric Administration (NOAA) convened a panel of prominent social scientists.[5] The product of the Panel's deliberations was a report that influenced the proposed NOAA regulations for natural resource damage assessment under the Oil Pollution Act of 1990 (published on January 7, 1994). The Panel's report concluded that:

> . . . under those conditions (and others specified above), CV studies convey useful information. We think it is fair to describe such information as reliable by the standards that seem to be implicit in similar contexts, like market analysis for new and innovative products and the assessment of other damages normally allowed in court proceedings.[6]

The Panel's "conditions" are a set of guidelines for CV survey design, administration, and data analysis. This paper focuses on one of these guidelines -- the Panel's call for the "temporal averaging" of willingness-to-pay (WTP) responses obtained from CV surveys as one method for increasing their reliability.[7] The Panel suggested:

> Time dependent measurement noise should be reduced by averaging across independently drawn samples taken at different points in time. A clear and substantial time trend in the responses would cast doubt on the "reliability" of the finding.

The NOAA Panel did not offer a clear description of the reasoning underlying its hypothesis that temporal averaging would enhance the reliability of CV estimates.[8] However,

---

[5] The panel was co-chaired by two Nobel Laureate economists, Kenneth Arrow and Robert Solow. Remaining members of the panel were: Edward Leamer of the University of California, Los Angeles, Paul Portney of Resources for the Future, Roy Radner of Bell Laboratories, and Howard Schuman of the University of Michigan. The panel's report was published in the January 15, 1993 issue of the *Federal Register.*

[6] *Federal Register*, January 15, 1993, p. 4610.

[7] Reliability as defined by the survey research community is a function of: (1) the underlying variation of the measure across the population of interest, (2) the concepts, wording, and method of presentation used in the survey, and (3) the nature of the sample used to make population inferences.

[8] In addition to temporal averaging, the Panel also recommended: (a) the use of probability samples allowing inference to target population, (b) personal interviews, (c) careful pretesting for interviewer effects and questionnaire design, and (d) the minimization of nonresponse. The panel also made specific

our findings suggest this proposal may be unnecessary, making their reasoning moot. Our findings, using a CV instrument designed to measure willingness-to-pay for a program to protect Prince William Sound, Alaska from future oil spills, like the Exxon Valdez spill, exhibited no significant sensitivity to the timing of the interviews. For two samples involving independent interviews taken over two years apart, the distribution of respondents' choices "for" and "against" the protection program did not differ.[9]

2. TESTING THE TEMPORAL VOLATILITY HYPOTHESIS

On March 24, 1989, the oil tanker Exxon Valdez left the port of Valdez, on its way to the Gulf of Alaska. It ran into the submerged rocks of Bligh Reef, rupturing its oil carrying compartments, and releasing some 11 million gallons of Prudoe Bay crude oil into the waters of Prince William Sound. As part of its damage assessment, the State of Alaska funded a CV study [Carson, *et al.* 1992] designed to measure the passive use losses due to the spill. With few exceptions, that study followed survey design and administration procedures identical to those subsequently recommended by the NOAA CV Panel. Our examination of the temporal averaging recommendation compares the results of the original national face-to-face survey conducted from January to mid-April 1991 with those of a follow-up, face-to-face survey

recommendations for the survey itself. These recommendations included: (a) a conservative survey design (*i.e.*, one that tends to understate values), (b) a willingness-to-pay referendum style value elicitation format, (c) accurate description of the program or policy, (d) pretesting of photographs, (e) reminder of undamaged substitute commodities, (f) adequate time lapse from the accident, (g) no-answer option, (h) yes/no follow-ups, and (i) checks on understanding and acceptance of the object of choice presented in the CV survey.

[9] Carson and Mitchell [1993] also report the results of a replication study. Their study, using a CV instrument to value changes in surface water quality, showed no significant differences in estimates of willingness-to-pay (after adjusting by changes in the consumer price index) between two surveys conducted three years apart.

conducted two years later using the identical questionnaire and a comparable sample. Because of the complexity of each study and the importance of the design and survey administration to the issue of reliability, we discuss each study separately.

After four field pilot tests, the original Exxon Valdez damage assessment survey was placed into the field in January of 1991, 22 months after the spill.[10] The field administration of the survey was conducted by Westat of Rockville, Maryland using a multi-stage area probability sample of residential dwelling units (DU) drawn from the United States and the District of Columbia. In the first stage of selection, 61 counties or county groups were drawn. Within these selected counties, about 330 blocks (or block groups) were chosen. In the third stage, approximately 1,600 dwelling units were drawn from the selected blocks.

The 61 first-stage selections consisted of Westat's National Master Sample of 60 PSUs (primary sampling units) which were drawn from the continental United States and the Honolulu SMSA which was drawn from the states of Alaska and Hawaii.[11] Westat's Master Sample of 60 PSUs was selected from a list that grouped the 3,111 counties and cities in the continental United States in 1980 into 1,179 PSUs, each consisting of one or more adjacent counties.[12]

Within each of the 61 PSUs, the second-stage selections were drawn from a list of all the Census blocks in the PSU. The lists were stratified by two block characteristics: percent of the population that was black, and a weighted average of the value of owner-occupied housing and the

[10] A complete description of the final survey and its development is provided in Carson *et al.* [1992].

[11] Because Alaska and Hawaii were excluded from Westat's original sampling list, a new stratum was created consisting of those two states. A random selection of PSUs from this stratum yielded the Honolulu SMSA.

[12] The 1980 census was used since results from the 1990 census were not available at the time the sample was drawn.

rent of renter-occupied housing. The 334 secondary selections were then drawn with probabilities proportionate to their total population counts. The overall response rate for the original study was 75.2 percent, yielding a sample of 1,043 cases.[13]

The estimates of lost passive use value due to the Exxon Valdez spill are reported in Carson *et al.* [1992]. These estimates provide a unique opportunity to examine whether CV studies of lost passive use conducted with survey protocols that meet most or all of the NOAA CV Panel guidelines yield the same WTP measures when the surveys are administered at two different points in time. To undertake this study, we administered the original Alaska survey questionnaire and visuals to a second sample of 300 US households about two years later. This second survey was conducted by the National Opinion Research Center (NORC) of the University of Chicago as part of a larger empirical study involving 1,408 interviewed households. The remaining 1,108 households received versions of the original Alaska instrument that were modified to examine other issues not relevant to this study.[14]

The second study was conducted in 12 PSUs selected from NORC's master area probability sample: Baltimore, MD; Birmingham, AL; Boston, MA; Charleston, SC; Harrisburg, PA; Ft. Wayne, IN; Manchester, NY; Nicholas County, KY; Portland, OR; Richmond, VA; Seattle, WA; and Tampa, FL. Six segments were selected from each PSU, resulting in 72 segments. Given past vacancy rates, 1,925 dwelling units were then randomly selected from the 72 segments. NORC's sampling staff then randomly assigned an interview version to each selected dwelling unit in advance of the field period.

[13] Non-English speaking households were ineligible for the survey.

[14] Results of the larger study are contained in Carson, Hanemann, *et al.*, 1994.

The selection of the respondent for the interview was made from all individuals in the household meeting the eligibility requirements: household member 18 years of age or older who owns, rents, or pays toward the mortgage or rent of the household.[15] The interviews for this study were conducted over an eight-week period from May 26 to July 17, 1993 and the overall response rate was 73 percent. As in the original survey, non-English speaking households were ineligible for the survey.

Due to differences in how PSUs were drawn in the first stage of sample selection, the original 1991 sample and the 1993 sample are not fully equivalent. In the 1991 sample, the first stage PSU selection followed a full probability selection scheme. That is, each of the 61 PSUs used in the sample were randomly selected from Westat's master sampling list. In the case of the 1993 sample, the 12 PSUs were selected from NORC's master list by choosing PSUs where NORC had sufficient interviewers to conduct the study. In all subsequent stages of sample selection (i.e., choosing Census blocks, dwelling units, and respondents), the samples were drawn identically. The effect of the difference in the first stage sampling was to exclude the major metropolitan areas of New York, Philadelphia, Chicago and Los Angeles (included in the 1991 sample) from the 1993 sample.

Since the first stage sampling differs in the 1991 and 1993 samples, we provide two different procedures to adjust for sample differences. First, we present in section four a set of results based on a functional specification designed for the 1991 sample that tests for differences in WTP while controlling for demographic differences in the respondents. Second,

[15] In households with more than one eligible respondent, the interviewer used a random number table to select one respondent for the main interview.

we conducted all of the analyses reported in this paper using a sub-sample of the of the 1991 sample that excluded the following PSUs: Bronx/Manhattan, NY; Kings/Queens/Richmond, NY; Nassau/Suffolk, NY; Philadelphia, PA; Chicago, IL; Los Angeles, CA. None of the test outcomes are changed when using this sub-sample.

3. RESULTS

The questionnaire used in the original 1991 survey and the 1993 replication employed a referendum, or discrete choice, value elicitation format. Respondents were asked to vote on a program that, for the next ten years, would protect Prince William Sound from another oil spill causing natural resource injuries comparable to those from the Exxon Valdez spill. In this framework, an initial question asks how the respondent would vote on the protection program if it cost their household $____. If the respondent said "for," she was asked in a follow-up question -- how would she vote if the program cost a higher amount? If the respondent answered "against" or "not sure" to the first question, the respondent was asked how she would vote if the program cost a lower amount. Four versions of the survey questionnaire, differing only in the amounts used in these two questions, were administered in-person to the sample.[16]

[16] The four versions of the original and follow-up surveys differed by the dollar amounts households were told they would pay in higher taxes if the prevention plan was adopted. The actual amounts used are displayed below.

Version	First Amount	Second Amount if "for" the program	Second Amount if "against" the program
A	$10	$30	$5
B	$30	$60	$10
C	$60	$120	$30
D	$120	$250	$60

There are at least three possible tests of the temporal averaging/reliability hypothesis based on the data from the two surveys. The first would be a simple test for differences in the pattern of choices ("for" and "against" votes) from the referenda questions across the two independent samples interviewed two years apart.

A second approach compares choice functions used to evaluate construct validity[17] based on the two surveys and tests for differences in the coefficients of variables hypothesized to be important to individuals' decisions. A choice function is a statistical model relating respondents' choices to their characteristics and other hypothesized casual variables. In the simplest case of open-ended questions, the respondent's willingness-to-pay (WTP) is regressed on respondent characteristics such as income, demographic variables, and attitude variables hypothesized to be relevant to the item being valued. If CV estimates of passive use losses are volatile over time, then the statistical relationships explaining these choices (or WTP responses) would also be expected to be different using independent samples taken at different times. With discrete choice questions, these models focus on the reported choices and use probit (or another maximum likelihood estimator for categorical variables) to describe the factors influencing those choices.

The third test, which is particularly relevant to the referendum or discrete choice format, considers the willingness-to-pay distributions based on CV choices estimated separately using the two surveys. The presence of volatility would be expected to imply significantly different willingness-to-pay estimates.

[17] Construct validity refers to the degree to which a measure relates to other measures predicted by theory. As a rule, two forms of construct validity are considered: convergent validity and theoretical validity. The former refers to whether the measure of interest is correlated with other measures of the same theoretical construct and is not applicable to the research proposed here. For an example of a test using contingent valuation, see Carson *et al.*, 1992.

Results for Contingency Table

Table 1 displays the percentage of respondents voting "for" or "against" adoption of the protection program based on the first question.[18] The table displays the percentages for the two surveys, for each of the four dollar amounts used. Casual examination of the distributions in Table 1 suggests little, if any, difference in response patterns between the two surveys. A chi-square test confirms that there is no evidence of time dependency in the responses.

Table 1: Percent Voting For/Against the Prevention Plan

	Response			
	For		Against	
Dollar Amount	1991	1993	1991	1993
$10	67	68	33	32
$30	52	55	48	45
$60	51	49	49	51
$120	35	33	65	67

Note: Both the 1991 and 1993 surveys permit respondents to reconsider their votes later in the survey. This analysis considers only the response to the first vote question and therefore does not reflect reconsideration of the vote.

More specifically, for each dollar amount asked ($10, $30, $60, $120), a two by two contingency table ("for"/"against" by 1991/1993) was constructed. The first column of **Table 2** presents the results for these tests of the null hypothesis that the frequencies of "for" and "against" the plan are the same for both the two surveys at each dollar amount.

[18] Throughout the analysis reported in the paper "don't know" or "not sure" responses to the voting questions are treated as "against" votes.

Table 2: 2 x 2 Contingency Table Results

	x^2 Statistics	
Dollar Amount	First Vote	First and Second Vote
$10	0.0176	0.9610
$30	0.0384	9.3095*
$60	0.2055	1.6360
$120	0.0193	0.4837

* Significant at the 95% confidence level

Note: Both the 1991 and 1993 surveys permit respondents to reconsider their votes later in the survey. In this analysis the column labeled "First and Second Vote" bases the outcome of the second vote on any reconsiderations the respondent made, that is, changing their vote from "for" to "against."

The second column of Table 2 presents the contingency table results using choices from the first and second voting questions. There are four possible voting patterns based on both vote questions -- for-for, for-against, against-for, and against-against. This voting pattern yields four 4 by 2 tables where the null hypothesis is tested using a χ^2 statistic with three degrees of freedom. Table 2 reveals that the null hypothesis of equal distribution can be rejected only at the $30 amount when the relevant voting pattern is based on both voting questions.

Results for Choice Function

Three choice function estimators have been used to test for consistency in the factors influencing respondents' choices in the two samples. The first is a probit model using the responses to the first referendum question. The second is a survival model with a Weibull specification based on the first question. The third also adopts the Weibull survival framework but, in this case, the responses to both the first and second voting questions are used in constructing the interval estimates.

Each of the estimators has quite different implicit assumptions. While both the probit and the first of the Weibull hazard (or survival) estimators rely on the responses to the first referendum question, they maintain different distributional assumptions. The Weibull is equivalent to assuming a model specified in terms of the log (WTP) that constrains the probability to vote "for" the program to be unity when the proposed tax amount is zero. The probit was estimated in terms of the level of the tax amount (and thus is consistent with a linear WTP specification). It does not constrain the probability of favoring the program as the tax amount declines to zero.

The double-bounded estimate is perhaps the most controversial estimator in that it relies on the responses to both questions being governed by the same underlying probability distribution over tax amounts. Under this assumption, the two questions provide greater resolution to the interval estimate of log (WTP). Cameron and Quiggin [1994] have suggested violations in this assumption can have important effects on the properties of the estimates of WTP and of the choice function.[19] We consider all three approaches in evaluating the properties of these estimated choice functions.

Table 3 defines the independent variables included in all choice models and corresponds to the regressors selected for the original 1991 survey [see Carson *et al.*, 1992, for a more complete discussion]. Because this analysis seeks to evaluate whether replication would change conclusions about choices, we did not consider alternative specifications.

[19] There have been a variety of responses to the critique. Kanninen [1995] argues implicitly that the bias could be due to poor bid design. Alberini's [1995a] analysis of the properties of different bid designs also "accepts" the responses to the second question as arising from the same underlying distribution as the first. In recent research, Alberini [1995b] has demonstrated the bias/variance tradeoff implied by assuming the responses to the two questions are perfectly correlated when in fact they are highly correlated would generally favor use of the double bounded estimator.

Table 3: Definition of Variables

Variable Name	Coding of Variable
constant	Intercept, equals unity for all respondents
1993	Coded as 1 if respondent from the 1993 replication sample, 0 otherwise
wlamt	dollar amount for first stated tax amount
linc	logarithm of household income
protest	response coded as 1 if respondent protested that Exxon or the oil companies should pay for the plan before they were asked how they would vote, 0 otherwise.
gmore	response coded as 1 if respondent answered B-1 as more damage and B-2 as 3 indicating great deal more damage than Exxon Valdez in absence of escort ship plan; 0 otherwise
more	response coded as 1 if respondent answered B-1 as more damage and B-2 as 2 indicating somewhat more damage than Exxon Valdez in absence of escort ship plan; 0 otherwise
less	response coded as 1 if respondent answered B-1 as less damage and B-3 as a little or a lot less than Exxon Valdez in absence of escort ship plan; 0 otherwise
nodam	response coded as 1 if respondent answered B-1 as less damage and B-3 as no damage in relation to Exxon Valdez in absence of escort ship plan; 0 otherwise
mwork	response coded as 1 if respondent answered plan not completely effective (B-7) and suggest in B-8 it would reduce damage a little or a moderate amount; 0 otherwise
nwork	responses coded as 1 if respondent answered plan not completely effective (B-7) and suggest in B-8 it would not reduce damage at all; 0 otherwise.
name	responses coded as 1 if respondent spontaneously named the Exxon Valdez as one of the major environment accidents caused by humans; 0 otherwise
coastal	response coded as 1 if respondent rated as personally (A-3) protecting coastal areas from oil spills as "extremely important" or "very important"; 0 otherwise.
wild	responses coded as 1 if respondent indicated (A-4) government should over next few years set aside very large amount or large amount of new land as wilderness; 0 otherwise.
sten	responses coded as 1 if respondent identifies himself (or herself) as a strong environmentalist (B-17 = 1 or 2); 0 otherwise.
likvis	response coded as 1 if respondent indicates household "very likely" or "somewhat likely" to visit Alaska in future; 0 otherwise.
white	response coded 1 for Caucasian, 0 otherwise.

Table 4: Choice Functions

Variable	Probit.		First Vote Survival		First & Second Vote Survival	
1993	–.011	(.097)	–.025	(.225)	.009	(.131)
w1amt	–.009*	(.001)	-	-	-	-
linc	.080	(.050)	.218	(.120)	.229*	(.068)
protest	–.944*	(.113)	–2.047*	(.304)	–1.169*	(.145)
gmore	.570*	(.160)	1.714*	(.515)	.759*	(.228)
more	–.693	(.960)	–1.671	(2.177)	.065	(1.494)
less	–.382*	(.099)	–.851*	(.235)	–.580*	(.129)
nodam	–.366	(.300)	–.882	(.595)	–.433	(.363)
mwork	–.069	(.084)	–.138	(.198)	–.203	(.113)
nwork	–1.400*	(.403)	–2.604*	(.694)	–1.848*	(.393)
name	.152	(.086)	.301	(.203)	.306	(.116)
coastal	.288*	(.107)	.485*	(.244)	.201	(.139)
wild	.154	(.086)	.403*	(.201)	.335*	(.114)
stenv	.135	(.091)	.362	(.220)	.297*	(.125)
likvis	.212*	(.090)	.519*	(.222)	.247*	(.123)
white	.320*	(.105)	.701*	(.247)	.287*	(.138)
_cons	–.731	(.504)	1.478	(1.193)	1.353*	(.673)

Notes: n = 1144
* indicates significance at the 95% level

Table 4 presents the probit and survival estimates. The data from the two surveys are pooled for a total of 1,144 observations.[20] The first column of Table 4 presents the probit results. Standard errors are shown in parentheses. The variable 1993, identifying the replication sample as an intercept shift, is insignificant. Thus, under the assumption of

[20] The original 1991 and the recent 1993 data sets employed in the contingency table tests had 1,043 and 300 observations respectively for a total of 1,343 observations. In the choice function equations we employ the logarithm of income as an explanatory variable. In the 1991 and 1993 data there are 160 and 39 observations respectively that have missing income information. This reduces the size of the pooled data set that can be used to estimate the choice functions to 1,144 observations.

common slope parameters, the probit equation model suggests that both the 1991 and 1993 choice functions have the same intercept.

The second and third columns of Table 4 offer the same conclusion, using the single and double-bounded Weibull survival models. Both imply that the 1993 coefficient is insignificant. Thus, from the analysis of the choice functions under the maintained assumption of identical slope parameters, there is no statistical difference between the intercept of the 1993 choice function and the intercept of the 1991 function.

Table 5 presents the results of relaxing the common slope parameter assumption. Each of the three models presented (probit, single-bounded survival, and double-bounded survival) contain a 1993 intercept shifting variable and interaction dummy variables (denoted N variable) for each of the slope parameters. The interactive dummy variables take on the value of the variable when the observation is drawn from the 1993 data set and a zero when the observation comes from the 1991 data.

The 1993 intercept shifting variable is again insignificant in all three models under the relaxed common slope parameter assumption. In the probit model, only the COASTAL interaction slope effect is significant at the five percent level. This is consistent with the single-bounded survival model where only the COASTAL slope is statistically significant. The double-bounded model suggests that *no* slopes are significantly different between the two choice functions. These results suggest that the overall choice determinants and their marginal effects as captured by the choice function estimates remained quite stable.

Table 5: Choice Functions With Full Interaction Effects

Variable	Probit.		First Vote Survival		First & Second Vote Survival	
1993	.905	(1.174)	1.300	(2.645)	1.371	(1.528)
w1amt	–.009*	(.001)	--	--	--	--
linc	.094	(.059)	.251	(.141)	.257*	(.078)
protest	–1.073*	(.135)	–2.292*	(.348)	–1.279*	(.166)
gmore	.591*	(.188)	1.778*	(.593)	.629*	(.252)
more	–.720	(.956)	–1.699	(2.138)	.014	(1.479)
less	–.319*	(.115)	–.648*	(.263)	–.453*	(.148)
nodam	–.451	(.366)	–1.184	(.716)	–.735	(.415)
mwork	–.147	(.097)	–.302	(.226)	–.274*	(.127)
nwork	–1.318*	(.407)	–2.425*	(.694)	–1.768*	(.393)
name	.141	(.100)	.229	(.233)	.253	(.131)
coastal	.435*	(.126)	.773*	(.287)	.358*	(.159)
wild	.095	(.098)	.244	(.229)	.269*	(.129)
stenv	.236*	(.106)	.575*	(.265)	.386*	(.146)
likvis	.146	(.105)	.335	(.252)	.181	(.143)
white	.342*	(.118)	.791*	(.278)	.335*	(.154)
n_w1amt	–.001	(.002)	--	--	--	--
n_linc	–.065	(.118)	–.103	(.268)	–.106	(.156)
n_prtest	.484	(.255)	1.003	(.547)	.463	(.330)
n_gmore	–.130	(.367)	–.377	(1.043)	.476	(.562)
n_less	–.330	(.237)	–.821	(.509)	–.552	(.300)
n_nodam	.342	(.659)	1.342	(1.354)	1.477	(.912)
n_mwork	.331	(.202)	.737	(.464)	.322	(.270)
n_name	.043	(.206)	.221	(.472)	.230	(.276)
n_coast	–.555*	(.249)	–1.129*	(.561)	–.595	(.325)
n_wild	.280	(.207)	.640	(.482)	.303	(.277)
n_stenv	–.384	(.216)	–.781	(.508)	–.455	(.293)
n_likvis	.252	(.211)	.653	(.507)	.179	(.285)
n_white	–.094	(.260)	–.344	(.585)	–.254	(.351)
_cons	–.937	(.591)	1.044	(1.405)	1.004	(.781)

Notes: n = 1144

* indicates significance at the 95% level

Willingness-to-Pay Estimates

The summary WTP statistic we have estimated from the two surveys is based on the Turnbull (1976) non-parametric, maximum likelihood (ML) estimator for interval-censored data. The Turnbull estimator uses respondents' choices to the voting questions to estimate the latent willingness-to-pay implied by each respondent's choice (*i.e.*, vote). As noted above, an individual's answer to a single question will distinguish either a lower or upper bound for his or her WTP. By combining respondents' choices, we obtain estimates for the relative frequency of responses at different WTP intervals, $(0, W1AMT_i)$ and $(W1AMT_i, \infty)$, where $W1AMT_i$ is one of the four tax amounts administered to the different sub-samples. The first pair, $(0, W1AMT_i)$, defines the interval identified by $W1AMT_i$ as an upper bound and, the second pair, $(W1AMT_i, \infty)$, with $W1AMT_i$ as a lower bound. The five intervals defined by W1AMT are: (1) \$0 to \$10, (2) \$10 to \$30, (3) \$30 to \$60, (4) \$60 to \$120, and (5) above \$120.

Two summary statistics can be defined based on the Turnbull estimates of the fraction of the sample in each of the five intervals. The first of these is the lower-bound mean. It is calculated using the fraction of the sample estimated to be in each interval to weight the lower end-point of the interval.[21] The second of these summary statistics is the upper-bound mean which is calculated in a similar manner by weighting the upper end of each interval by the fraction of respondents estimated to be in each interval.[22] The unobserved mean is always

[21] For instance, if 20% of the sample is estimated to be in the interval \$10 to \$25, the lower-bound mean is calculated by assuming that this 20% of the sample is willing to pay exactly \$10.

[22] The upper-bound mean is potentially infinite unless additional assumptions (such as no respondent would be willing to pay more than some fraction of his or her income) are imposed.

bounded from below by the estimated lower-bound mean and from above by the estimated upper-bound mean.[23]

The Turnbull lower-bound estimate of mean WTP from the original 1991 study, using responses to the first voting question, is $52.80 with a standard error of $2.12. The comparable estimate for the 1993 data is $52.81 with a standard error of $4.08. There is no significant difference between the two lower-bound means at reasonable levels of statistical significance.

In Carson, Mitchell, *et al.* [1995], the lower-bound mean is estimated using both the first and second vote questions. Employing both the first and second vote choices, and the reconsideration questions, yields the following seven WTP intervals: (1) $0 to $5, (2) $5 to $10, (3) $10 to $30, (4) $30 to $60, (5) $60 to $120, (6) $120 to $250, and (7) above $250. The lower bound Turnbull mean based on these seven intervals and using the 1991 data is $54.23 ($2.72), while the comparable estimate based on the 1993 data is $54.02 ($5.13). As with choices and choice functions, there is no statistical difference between the WTP measures.

4. CONCLUSION

The NOAA CV Panel recommended that estimates of WTP derived from CV studies be averaged over time to increase their reliability (validity) by reducing "time dependent measurement noise." The Panel also noted that a "clear and substantial time trend in the

[23] This statement is true irrespective of the particular amounts used to define the intervals, although the particular tax amounts used can influence how much *less* the lower-bound mean is than the sample mean and how much *greater* the upper-bound mean is than the sample mean. Random assignment of respondents to tax amounts will result in subsamples at each tax amount which are approximately equivalent in finite samples.

responses would cast doubt on the 'reliability' of the finding." This paper has sought to evaluate the importance of time induced volatility in estimates of WTP derived from contingent valuation by evaluating the potential for WTP volatility using a CV study that meets most of the Panel's recommendations and guidelines.

Three features of the stated choices of our respondents are examined that might vary over time (recall the respondents were faced with a choice regarding a vote "for" or "against" a program to protect Prince William Sound from another oil spill like the Exxon Valdez). These features are: (1) the distribution of "for" and "against" votes, (2) parameters of estimated choice functions, and (3) estimates of WTP. Choices were not significantly different. Two sets of estimates (a nonparametric lower- and upper-bound) for the mean WTP vary by no more than 25 cents over the 2-year period and the choice functions are remarkably stable.

Nonetheless, it is important to acknowledge that this is only one test of temporal volatility issue. Our findings concur with the temporal stability of valuation estimates displayed in earlier test/retest studies [see Loomis, 1989, 1990 as one example].[24] Taken together these studies suggest that the Panel's concerns are unsubstantiated and not as important as its recommendation could be interpreted to imply. Our example involved a large, exceptionally well-known incident where the media coverage alone might have been expected to influence people's choices. This is not what our results imply despite the passage of two years between the original and follow-up surveys and over four years from the time of the incident. Thus, these findings may suggest that when stated choices are elicited with CV

[24] An unpublished test/retest of Smith and Desvousges [1986] demand for distance model two years after the initial survey conducted by Smith and Gregory Michaels yielded virtually identical estimates of the price and income elasticities in their demand model.

questions that include detailed background information and that have been pretested, the effects of temporal volatility are generally not evident. At this point, given the evidence available, it appears that temporal volatility may not impact on CV's reliability for surveys that meet the NOAA Panel's general guidelines.

REFERENCES

Alberini, A. 1995a. "Willingness-to-Pay Models," *Land Economics*, vol. 71, no. 1, pp. 83-95.

Alberini, A. 1995b. "Efficiency vs. Bias of Willingness-to-Pay Estimates: Bivariate and Interval Data Models," *Journal of Environmental Economics and Management* (forthcoming).

Arrow, Kenneth, Robert Solow, Edward Leamer, Paul R. Portney, Roy Radner, and Howard Schuman. 1993. Report of National Oceanic and Atmospheric Administration panel on the reliability of natural resource damage estimates derived from contingent valuation, *Federal Register*, January 15.

Cameron, T. A., and J. Quiggin. 1994. "Estimation Using Contingent Valuation Data from a 'Dichotomous Choice with Follow-up' Questionnaire," *Journal of Environmental Economics and Management*, vol. 27, no. 3, pp. 218-234.

Carson, R. T., W. M. Hanemann, R. J. Kopp, J. A. Krosnick, R. C. Mitchell, S. Presser, P. A. Ruud and V. K. Smith. 1994. "Prospective Interim Lost Use Value Due to DDT and PCB Contamination in the Southern California Bight," Report to National Oceanic and Atmospheric Administration, Natural Resource Damage Assessment, Inc., La Jolla, Calif., September.

Carson, R. T. and R. C. Mitchell. 1993. "The Value of Clean Water: The Public's Willingness to Pay for Boatable, Fishable and Swimmable Quality Water," *Water Resources Research*, vol. 29, no. 7, pp. 2445-2454.

Carson, R. T., R. C. Mitchell, W. M. Hanemann, R. J. Kopp, S. Presser, and P. A. Ruud. 1992. "A Contingent Valuation Study of Lost Passive Use Values Resulting From The Exxon Valdez Oil Spill," Report to the Attorney General of the State of Alaska.

Carson, R. T., Mitchell, R. C., Hanemann, W. M., Kopp, R. J., Presser, S. and Ruud, P. A. 1995. "Contingent Valuation and Lost Passive Use: Damages From The Exxon Valdez," Discussion Paper, Department of Economics, University of California, San Diego.

Carson, R. T., J. Wright, N. Carson, A. Alberini, and N. Flores. 1995. "A Bibliography of Contingent Valuation Studies and Papers," La Jolla, Calif., Natural Resource Damage Assessment, Inc.

Diamond, P. and J. A. Hausman. 1994. "Contingent Valuation: Is Some Number Better than no Number," *Economic Perspectives*, vol. 8, no. 4, pp. 45-64.

Hanemann, W. M. 1994. "Contingent Valuation and Economics," *Economic Perspectives*, vol. 8, no. 4, pp. 19-44.

Hausman, J. A. 1993. *Contingent Valuation: A Critical Assessment* (Amsterdam, Elsevier Science Publishers B. V.).

Kanninen, B. J. 1995. "Bias in Discrete Response Contingent Valuation," *Journal of Environmental Economics and Management*, vol. 28, no. 1, pp. 114-125.

Kopp, R. J., and K. A. Pease. 1996. "Contingent Valuation: Economics, Law and Politics," in Kopp, Pommerehne and Schwarz (eds.), *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods* (Boston, Mass., Kluwer-Nijhoff), forthcoming.

Loomis, J. B. 1989. "Test-Retest Reliability of Contingent Valuation Method: A Comparison of General Population and Visitor Response," *American Journal of Agricultural Economics*, vol. 71, pp. 76-84.

Loomis, J. B. 1990. "Comparative Reliability of the Dichotomous Choice and Open-Ended Contingent Valuation Techniques," *Journal of Environmental Economics and Management*, vol. 18, pp. 78-85.

McClelland, G. H., W. D. Schulze, J. K Lazo, D. W. Waldman, J. K. Doyle, S. R. Elliott, and J. R. Erwin. 1992. "Methods for Measuring Nonuse Values: A Contingent Valuation Study of Groundwater Cleanup," Report to the U.S. Environmental Protection Agency.

Nelson, W. 1982. *Applied Life Analysis* (New York, John Wiley).

Portney, P. R. 1994. "The Contingent Valuation Debate: Why Economists Should Care," *Economic Perspectives*, vol. 8, no. 4, pp. 3-18.

Smith, V. K., and W. H. Desvousges. 1986. "The Value of Avoiding a LULU: Hazardous Waste Disposal Sites," *Review of Economics and Statistics*, vol. 68, pp. 293-99.

Turnbull, B. W. 1976. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society*, B38, pp. 290-295.