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 panel recommended that "time dependent measurement noise should be reduced by averaging across independently drawn samples taken at different points in time." In this paper we 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, exhibited no significant sensitivity to the timing of the interviews.* (JEL Q26, D60)

I. 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 studies in the literature dealing with CV (see Carson et al. 1995). Special attention has focused on its use in estimating passive use value because indirect methods can only measure use-related values.[1] While there is a substantial literature describing the theoretical foundation for nonuse or passive use values (e.g., Krutilla 1967; Plourde 1975; McConnell 1983), the prospect of routinely including estimates for these losses in natural resource damages has generated considerable controversy.[2] The 1989 Court of Appeals ruling in *Ohio v. Department of the Interior* held that lost passive use values should be included in damage awards resulting from injuries to natural resources due to releases

The authors are, respectively: associate professor of economics, University of California (San Diego); 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 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, Alan Randall, and two anonymous referees 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, and bequest value. Option value is also listed as a component of passive use value in some of the discussion explaining the ruling. The literature now generally recognizes option value as a measure of people's risk aversion for factors that might affect the ability to have access to or use environmental resources and therefore not a component of nonuse values (see Smith 1987 and Randall 1991). In addition, subsequent research by Larson (1993) has suggested that existence values could be measured with assumptions from information about people's use of the resource to be valued. While Larson's derivation is correct, his interpretation requires specific untestable assumptions restricting individual preferences to offer the interpretation as existence values (Bockstael and McConnell 1993).

[2] See Diamond and Hausman (1994) and Hanemann (1994) as examples.

of hazardous substances.[3] Under this decision, it is unnecessary for an individual to be a direct user of a natural resource, for example as a recreationist, to hold an economic value for the resource in question.[4] The *Ohio* Court also emphasized the importance of the "reliability" of the methods used to estimate natural resource damages.[5] Because contingent valuation is currently the only technique available to measure economic values that include use and passive use, much of the current CV research has been directed at evaluating its reliability.

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

The Panel's "conditions" are a set of guidelines for CV survey design, administration, and data analysis.[7] 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 findings. (*Federal Register*, January 15, 1993, 4609)

The reasoning underlying the NOAA Panel's recommendation for temporal averaging is not clear. Measurement error can be reduced by averaging across multiple observations that are assumed to be realizations from the same underlying stochastic

[3] 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.

[4] We adopt the term "economic value" rather than simply "value" to distinguish our meaning from other uses of the word value. Economic values are defined by an individual's choices. When it is known that someone chooses to give up x in order to obtain y, we can conclude the economic value of y (termed the object of choice) is at least x.

[5] 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 (1997), 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 in terms of *replicability*: the extent to which the same results are obtained when the identical measurement procedures are repeated. We use the term reliability in this latter sense.

[6] The Panel was co-chaired by two Nobel Laureate economists, Kenneth Arrow and Robert Solow. The additional 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 New York University, and Howard Schuman of the University of Michigan. The Panel's report was published in the January 15, 1993, issue of the *Federal Register*.

[7] 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 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

process. The Panel's report raises concern about the existence of time dependency in the location or scale parameters for CV constructed measures of willingness to pay.

However each individual's economic value for a commodity *should* be expected to change with the conditions that influence any choice. In general, the prices (and availability) of substitutes and complements, level of income, and all other factors that would affect these decisions can be expected to be determinants of measures of economic values.[8] Thus, changes in estimates of economic value, alone, are not likely to be the source of the NOAA Panel's call for attention to the temporal reliability of CV. Rather it might represent a concern that immediate reactions to an event, such as a large oil spill, may be particularly labile. Thus, for example, public reaction might initially entail outrage directed at the party thought to be responsible, or more generally, people may require time to evaluate the full implications of the event. With time, such short-term responses may be often modified as more information about the cause, and the full consequences of the event, becomes known. The Panel's suggestion might be treated as a concern over the timing of a single CV survey in relation to the event giving rise to natural resource injuries.

In this context temporal averaging would not improve the estimates. Their hypothesis implies WTP estimates constructed from one set of responses would be superior to those at the times that are more subject to these short-term influences. Given such concerns, it is important to distinguish research on the stability of CV estimates of WTP over time from a recommendation to average the estimates for increased reliability. The alternative hypotheses providing reasons for focusing research on temporal stability all suggest temporal averaging would not improve the properties of CV estimates.[9]

Our findings use a CV instrument designed to measure WTP for a program to protect Prince William Sound, Alaska, from future oil spills, like the Exxon Valdez spill. These results indicate that choices made two years after the spill are not significantly different from those made four years after the spill.[10]

II. 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, 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 the survey design and administration procedures subsequently recommended by the NOAA CV Panel. The Exxon Valdez spill together with the Carson et al. study offer a unique opportunity to investigate the question posed by the NOAA Panel. By conducting a comparable analysis four years after the spill we can investigate whether the timing of this initial study was within the interval the Panel implicitly suspected could be problematic. To undertake our analysis, we compare the results of the

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.

[8] In the case of resources that are assumed to provide a source of passive use values it is reasonable, following Hanemann (1988), to assume they make separable contributions to individual preferences. Under this premise we would expect that the amount and conditions of access to other substitute resources could influence these choices, but the relative prices of other goods making the separable contribution to preferences would not. In that case measures of economic value would respond to changes in only the aspects of the circumstances of choice related to income and nonmarket substitutes.

[9] There has been some evidence that news and its sources influence public opinions. See Jordan (1993) as one example.

[10] 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.

original national face-to-face survey conducted from January to mid-April 1991 with those of a follow-up, face-to-face survey conducted in 1993 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.[11] The field administration of the survey was conducted by Westat, one of the nation's leading survey organizations, using a multi-stage area probability sample of residential dwelling units (DU) drawn from the 50 United States and the District of Columbia. The Primary Sampling Units (PSUs) consisted of Westat's National Master Sample supplemented by the Honolulu SMSA.[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 rent of renter-occupied housing. The 334 secondary selections were then drawn with probabilities proportionate to their total population counts. In the third stage, approximately 1,600 dwelling units were drawn from the selected blocks. Within each dwelling unit, a household member 18 or older who owned, rented, or paid toward the mortgage or rent was selected at random to be the respondent. The overall response rate for the original study was 75.2 percent, yielding a sample of 1,043 cases.[13]

Our second survey was conducted by the National Opinion Research Center (NORC) of the University of Chicago as part of an empirical study involving 1,408 interviewed households. Three hundred of these respondents received the original Alaska questionnaire and visuals. 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 sample was composed of 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. 1,925 dwelling units were then randomly selected from the 72 segments. NORC's sampling staff then randomly assigned one of four interview versions of the questionnaires comprising our larger study to each selected dwelling unit in advance of the field period.

The selection of the respondent for the interview was made from all individuals in the household meeting the same eligibility requirements as with the original 1991 Exxon Valdez survey.[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. The 12 PSUs in the 1993 sample were selected from NORC's master list by choosing

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

[12] Westat's Master Sample of 60 PSUs was selected from a list that grouped the 3,111 counties in the continental United States in 1980 into 1,179 PSUs, each consisting of one or more adjacent counties. The 1980 census was used since results from the 1990 census were not available at the time the sample was drawn. 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.

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

[14] Results of the larger study are contained in Carson et al. (1994).

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

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 with identical procedures. One 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. Section IV presents results based on a choice function, conventionally used in tests of construct validity (Mitchell and Carson 1989). The specification for this function was based on the construct validity test with the 1991 sample. We use this function to test for differences in the parameters associated with the factors influencing choices with the two samples. In addition, we replicated all of the analyses reported in this paper using a subsample 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. Therefore, we focus our discussion on analyses that compare the full 1991 sample with our 1993 replication.

III. RESULTS

The questionnaire uses a referendum 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. Questions were also asked in a double-bounded format so that if the respondents said they voted "for" the protection program then they were asked about a higher one-time cost question. Respondents answering "against" or "not sure" to the first amount were offered the program at a lower amount. Four versions of the base survey questionnaire, differing only in the amounts used in these two questions, were administered.[16]

Tests for the effects of the timing of the initial Alaska survey were undertaken in three ways: simple contingency analyses with both the first and the second response; analysis of the estimated parameters for the choice functions from each sample; and estimates of the WTP from each sample. We consider each in turn.[17]

Results for Contingency Table

The first panel in Table 1 reports the percentage of respondents voting "for" or "against" adoption of the protection program based on the first question. The table displays the percentages for the two surveys, for each of the four dollar amounts used. Simple inspection of the distributions suggests that the results of the initial Alaska survey were not impacted by its proximity to the incident. The identical survey conducted two years later provides equivalent results.

The null hypothesis of equal proportions voting "for" and "against" the plan is tested

[16] 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

[17] An approximate way to consider the power of our test of reliability is to use the Mitchell-Carson (1989, 365–66) evaluation of sample requirements to isolate specified differences in means expressed in proportional terms. Given the p-value for probability of a Type I error for the test, the desired power, and an assumption about the coefficient of variation for the initial sample, Table C-4 provides (for a two-sided t-test) the desired sample size when $\alpha = .05$ and power $= .90$. Taking the assumed coefficient of variation (cv) as the estimate based on the lower bound mean for 1991 we have $cv = .04$. This is substantially below the estimates in Table C-4. Nonetheless, using $cv = .1$ as a conservative assumption, a sample size of 28 would be required to detect 10 percent differences in the means (32 for power = .95). Our sample of 300 clearly exceeds this standard.

TABLE 1
CONTINGENCY ANALYSIS OF VOTES FOR/AGAINST PREVENTION PLAN

First Dollar Amount	Percent Voting For/Against Plan[a]				Contingency Test-χ^2			
	For		Against		First Vote[b]			First and Second Vote[c]
	1991	1993	1991	1993	For/Against	With DK	Without DK	
$10	67	68	33	32	0.0176	0.0333	0.0007	0.9610
$30	52	56	48	41	0.0384	0.4694	0.4690	9.35095*
$60	51	49	49	51	0.2055	0.2480	0.0040	1.6360
$120	34	33	66	67	0.0193	0.4868	0.0000	0.4837

*Significantly different at the 95 percent level.
[a]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.
[b]"For/Against" recodes volunteered "don't know/not sure" responses as "against." With DK includes "for," "against," and "don't know/not sure" as separate categories. Without DK drops the "don't know/not sure" responses from the sample.
[c]"First and Second Vote" base the outcome of the second vote on any reconsiderations the respondent made, that is, changing their vote from "for" to "against."

four ways with these choices. Using the first question we consider: (a) votes with "don't know" and "not sure" recoded as against;[18] (b) "don't know" and "not sure" treated as a separate category so three responses are allowed (i.e., "for," "against," and "don't know/not sure"); and (c) deleting the "don't know/not sure" responses. The next three columns in Table 1 report the chi square statistics for each possible interpretation of the choices reported with each dollar amount. None would permit rejection of the null hypothesis of equal proportions in the categories identifying the respondents' choices.

The last column in Table 1 presents the results using choices from the first and second voting questions. There are four possible voting patterns based on both response questions—for-for, for-against, against-for, and against-against. The null hypothesis of equal distribution between the two surveys can be rejected only at the $30 amount.

Results for Choice Function

Three estimators for the choice function were used in testing consistency as part of construct validity tests for respondents' choices in the two samples. Both probit and Weibull survival models were applied to the responses from the first question. In addition, we used the responses to both questions to develop interval censored estimates of a WTP function (i.e., the so-called double-bounded model, see Hanemann, Loomis, and Kanninen 1991) and again used a Weibull framework to evaluate the factors influencing the choices used in estimating this equation.

Each of the estimators has quite different implicit assumptions. The probit was estimated in terms of the level of the tax amount (and thus is consistent with a linear random utility or WTP specification, see McConnell 1990). It does not constrain the probability of favoring the program to unity as the tax amount declines to zero. The Weibull's location parameter assumes a model that implies independent variables in linear form will shift the log of median (or mean) WTP. It also constrains the probability to vote "for" the program to be unity when the proposed tax amount is zero.

The double-bounded estimator is perhaps the most controversial approach in that it relies on the responses to both questions being governed by the same underlying probability distribution. Cameron and Quiggin (1994) have suggested violations in this assumption can bias the estimates of WTP

[18] These responses were not offered by interviewers but were recorded if respondents voluntarily offered either answer.

and of the parameters in the WTP function used to describe the choices.[19] Our primary concern here is with the consistency in the overall conclusions from fitting these models to both samples.

Table 2 defines the independent variables included in all choice models. These factors correspond 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. Table 3 presents the probit and survival function estimates. The data from the two surveys are pooled for a total of 1,144 observations.[20] The first column of Table 3 presents the probit results. Standard errors are shown in parentheses beside the coefficients. The variable labeled as "1993" identifies the replication sample as an intercept shift. It is not significantly different from zero, implying that under the assumption of common slope parameters, there is no shift in the choice model. The second and third columns of Table 3 imply the same conclusion, using the single- and double-bounded Weibull survival models.

Table 4 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 independent variables (to allow testing for differences in each parameter between the two samples). The 1993 intercept shifting variable is again insignificant in all three models. With the probit and single-bounded survival estimates, only the *COASTAL* interaction slope effect would be judged significantly different from zero at the 5 percent level. Estimates for the double-bounded model imply none of the slope parameters are significantly different between the two choice functions for the two samples. Overall, then, the determinants of choices in the samples separated by two years remained stable.[21]

Willingness-to-Pay Estimates

Our estimates for the mean WTP use the Turnbull (1976) nonparametric estimator based on interval censored data along with Carson et al.'s (1994) method for estimating a lower bound for the mean of the underlying WTP distribution. Assuming referendum questions with a single take-it or leave-it decision, the design of responses over proposed costs, t_j allows respondents to be sorted into two groups for each cost (or tax amount). This allows the distribution function to be defined as:

$$\Phi_j = \text{Probability}(\text{WTP} \leq t_j)$$

$$1 - \Phi_j = \text{Probability}(\text{WTP} > t_j).$$

To develop a maximum likelihood estimator for the distribution function we need only the frequencies in each cell. The log-likelihood function, l, is given in equation [1].

$$l = \sum_{j=1}^{k} [N_j \ln(\Phi_j) + Y_j \ln(1 - \Phi_j)] \qquad [1]$$

where

N_j = number of respondents indicating "against" program at tax amount t_j,

[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 (1995) 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.

[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.

[21] There are actually different hypotheses implied by each estimator. With the probit model, the parameters reflect both the location and scale parameters for the distribution. In the Weibull model, they measure the percentage change in latent WTP with a change in each independent variable.

TABLE 2
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	Response 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	Response 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	Response 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	Response 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

Y_j = number of respondents indicating "for" program at tax amount t_j,
k = number of values for t_j.

The lower-bound estimate of mean WTP is defined in equation [2].[22]

$$\begin{aligned} \text{WTP}_{LB} = {} & 0 \cdot \text{Prob}(0 \le \text{WTP} < t_1) \\ & + t_1 \cdot \text{Prob}(t_1 \le \text{WTP} < t_2) \\ & + t_2 \cdot \text{Prob}(t_2 \le \text{WTP} < t_3) \\ & + \ldots + t_{k-1} \\ & \quad \cdot \text{Prob}(t_{k-1} \le \text{WTP} < t_k) \\ & + t_k \cdot (1 - \Phi_k). \end{aligned} \qquad [2]$$

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

The Turnbull lower-bound mean estimate from the 1991 sample, using responses to the first voting question, is $52.80 with a standard error of $2.12. The comparable estimate for the 1993 sample is $52.81 with

[22] Estimation with two questions yields interval estimates of Φ (e.g., $(\Phi_j - \Phi_{j-1})$). The likelihood function can be defined using these intervals. See Haab and McConnell (1996) for further illustration of the method.

[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.

TABLE 3
CHOICE FUNCTIONS

Variable	Probit.		First Vote Survival		First & Second Vote Survival	
1993	−.011	(.097)	−.025	(.225)	.009	(.131)
wlamt	−.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)

Note: $n = 1{,}144$.
* Indicates significance at the 95 percent level.

a standard error of $4.08. Whether or not we adjust for the effects of changes in the general price level over this time, there is no significant difference between the two samples' lower-bound means.[24]

Moreover, as one would suspect from the tests using contingency tables, our conclusions are insensitive to the treatment of "don't know/not sure" responses. Deleting them from the sample yields a lower-bound mean of $56.41 (2.21) for the 1991 sample and $57.27 (4.33) for the 1993 sample, with an asymptotic Z statistic (0.89) indicating no significant difference.

Using the first and second vote choices and the reconsideration questions to construct interval censored measures for estimating the distribution functions yields 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 sample is $54.23 ($2.72), while the comparable estimate based on the 1993 sample is $54.02 ($5.13). As with the cross tabulations for choices alone and the choice functions, these estimates are not significantly different.

IV. CONCLUSION

Three features of the stated choices of our respondents that might vary over time have been examined. They are (1) the distribution of "for" and "against" votes, (2) parameters of estimated choice functions, and (3) lower-bound estimates for the mean WTP. Choices were not significantly different. Several sets of estimates for the lower-bound mean of WTP were not significantly different in real terms, and the choice functions were remarkably stable.

We *should* expect estimates of the WTP for any object of choice to change as important aspects of the circumstances of choice change. The NOAA Panel's recommendation to consider evidence of "a clear and substantial time trend in responses" as a source of "doubt on the 'reliability' of the findings" is best interpreted as a concern about the timing of CV surveys in relation-

[24] Using the consumer price index to adjust for the price increases scales the 1991 estimate by 1.061. Then the asymptotic Z-statistic testing equality of the two means is 1.16, implying the null hypothesis of equality cannot be rejected at any conventional p-value.

TABLE 4
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)

Note: $n = 1{,}144$.
* Indicates significance at the 95 percent level.

ship to the date of the accident that may have prompted interest in measuring passive use losses (i.e., for a damage assessment). Our results suggest that a random sample of respondents' choices four years after the Exxon Valdez accident do *not* imply economic values that would be judged to be significantly different from what an independent sample selected in 1991 stated.

These results are remarkably stable and have prompted some diverse responses. For example, in contrast to the NOAA Panel's concerns about too much change, one might ask is there too little change?[25] Proponents of these questions might cite the apparent decline in the percent of people identifying environmental issues as problems that most concerned them (as reported from surveys by Roper Starch Worldwide, Inc.), as well as the decline in the percent reporting that we are spending too little on improving and protecting the environment (as reported from surveys by the National Opinion Research Center) over the approximate period covered by Exxon Valdez (1991) and NORC replication (1993) surveys.[26] To believe that changes in these broad indicators of environmental attitudes should be reflected in CV measures of WTP one must assume that CV responses are dominated by broad environmental attitudes rather than preferences

[25] We are grateful to John Payne for identifying this interpretation of the results.

[26] See Ladd and Bowman 1995.

for the specific plan to protect Prince William Sound. If this assumption is valid, one should not expect to see strong relationships between features of the plan and WTP. In this study we find such strong relationships.

Others might argue that the incomes and prices faced by households changed between 1991 and 1993 and therefore one should have expected more variability in estimates of WTP. For these concerns to be meaningful we need to be more specific about how these types of changes would be expected to influence measures of passive use values.

Consider first arguments that general price inflation or changes in the availability of market goods should have changed respondents' choices more directly than what we observe. If respondents' choices are motivated by concerns that would lead to passive use values, then *by definition* they are *not* linked to changes in the prices or availability of market goods. This follows because marketed goods must be assumed to make separable contributions to individual well-being from the environmental resources associated with the passive use value. This condition is implied by the definition of passive use (nonuse) values. As a result changes in the relative prices of marketed goods are unlikely to influence people's decisions for these types of environmental resources.[27]

Changes in income could influence monetary measures of passive use value. To evaluate the importance of this effect for our samples we considered respondents' reported household (before tax) income for 1990 and 1992 (for the 1991 sample and the 1993 replication). By converting the two estimates of the mean household income to 1993 dollars (using the CPI) we can compare the importance of income changes for stated choices and WTP estimates. The means are $37,231 for 1991 and $39,953 in 1993.[28] The lower-bound mean WTP in 1993 dollars was $56.02 for the 1991 sample and $52.81 for 1993. Thus, this type of fairly simple comparison offers little to suggest lower-bound mean estimates for WTP are inconsistent with the changes we would expect based on the absence of important changes in the economic circumstances of the households in 1991 and 1993.

In interpreting our findings it is important to acknowledge that this is only one test of temporal stability. Our findings do concur with the earlier test/retest studies (see Loomis 1989 as one example). Taken together with these studies they seem to suggest that the Panel's concerns about temporal stability may not be as important an issue as the Panel's overall recommendation might 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. Of course, we do not know what the pattern of responses would have been had the original survey been con-

[27] To the extent it is possible to isolate substitution relationships with other nonmarket resources, we might also expect that changes in these substitute resources would also influence measures of WTP. Nonetheless, the magnitude of these responses cannot be predicted *a priori*. At best, we have limited overall expectations from economic theory about changes in measures of WTP with changes in each individual's circumstances of choice.

[28] The comparison of average incomes leads to a bit smaller discrepancy if we focus on the full 1993 sample. The mean in this case is $38,305 (in 1993 dollars). A larger difference in income arises with a different treatment of the right censored highest category of income. Using the U.S. Department of Commerce (1966) approach to fitting a Pareto tail to the distribution, our adjusted (to 1993) mean income levels become 39,410 for the 1991 sample and 43,125 for the 1993 sample. This large discrepancy arises because a greater number of respondents in the 1993 sample reported incomes in the highest two income classes, and thus the mid-point assigned to the right censored class was greater. In the other computations the same censoring point was assigned to both samples.

To evaluate whether the choices were consistent with this income increase between the 1991 and 1993 surveys we computed chi square tests for each initial tax amount for the income group in this greatest income class. Only in the case of the $120 tax amount did we find significantly different choice patterns between the 1991 and 1993 samples ($p = .023$). A higher fraction of the 1993 sample supported the plan at this cost than the 1991 sample. This is consistent with what we would expect with the higher income levels. This result should be interpreted cautiously because only 12 respondents with this income level in the 1993 sample were assigned to this tax amount.

ducted closer to the time of the Exxon Valdez oil spill. As a result, our findings do not answer the fundamental question about when CV surveys should be conducted in relation to the timing of large, potentially controversial events like the Exxon Valdez spill. We can say that longer term averaging or trend analysis seems unwarranted in tests of the reliability of CV surveys.

References

Alberini, A. 1995. "Testing Willingness-to-Pay Models of Discrete Choice Contingent Valuation Survey Data." *Land Economics* 71 (Feb.):83–95.

Bockstael, N. B., and K. E. McConnell. 1993. "Public Goods as Characteristics of Nonmarket Commodities." *Economic Journal* 103 (9):1244–57.

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* 27 (3):218–34.

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, CA. 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* 29 (7):2445–54.

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

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

Haab, T. C., and K. E. McConnell. 1996. "Count Data Models and Recreational Demand." *American Journal of Agricultural Economics* 78 (1):89–102.

Hanemann, W. M. 1988. "Three Approaches to Defining 'Existence' or Nonuse Values Under Uncertainty." Department of Agricultural and Resource Economics, University of California, Berkeley. July.

———. 1994. "Contingent Valuation and Economics." *Journal of Economic Perspectives* 8 (4):19–44.

Hanemann, W. M., J. B. Loomis, and B. J. Kanninen. 1991. "Statistical Efficiency of Double-Bounded Dichotomous Choice Contingent Valuation." *American Journal of Agricultural Economics* 72 (4):1255–63.

Jordan, Donald L. 1993. "Newspaper Effects on Policy Preferences." *Public Opinion Quarterly* 57:191–204.

Kanninen, B. J. 1995. "Bias in Discrete Response Contingent Valuation." *Journal of Environmental Economics and Management* 28 (1):114–25.

Kopp, R. J., and K. A. Pease. 1997. "Contingent Valuation: Economics, Law and Politics." In *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods*, eds. R. J. Kopp, W. Pommerehne, and N. Schwarz. Boston: Kluwer-Nijhoff (forthcoming).

Krutilla, J. V. 1967. "Conversation Reconsidered." *American Economic Review* 57 (4):777–86.

Ladd, Everett Carl, and Karlyn H. Bowman. 1995. *Attitudes Toward the Environment: Twenty-five Years After Earth Day*. Washington, DC: The American Enterprise Press.

Larson, Douglas. 1993. "On Measuring Existence Value." *Land Economics* 69 (Nov.):377–89.

Loomis, J. B. 1989. "Test-Retest Reliability of the Contingent Valuation Method: A Comparison of General Population and Visitor Responses." *American Journal of Agricultural Economics* 71 (1):76–84.

———. 1990. "Comparative Reliability of the Dichotomous Choice and Open-Ended Contingent Valuation Techniques." *Journal of Environmental Economics and Management* 18 (1):78–85.

McConnell, K. E. 1983. "Existence and Bequest Value." In *Managing Air Quality and Scenic Resources at the National Parks and Wilderness Areas*, eds. R. D. Rowe and L. G. Chestnut. Boulder, CO: Westview Press.

———. 1990. "Models for Referendum Data: The Structure of Discrete Choice Models for Contingent Valuation." *Journal of Environmental Economics and Management* 78 (1):19–34.

Mitchell, Robert C., and Richard T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method.* Washington, DC: Resources for the Future.

Plourde, C. 1975. "Conservation of Extinguishable Species." *Natural Resources Journal* 15 (4):791–97.

Randall, Alan. 1991. "Nonuse Benefits." In *Measuring the Demand for Environmental Commodities*, eds. J. B. Braden and C. D. Kolstad. Amsterdam: North Holland.

Smith, V. Kerry. 1987. "Nonuse Values in Benefit Cost Analysis." *Southern Economic Journal* 54 (July):19–26.

Turnbull, B. W. 1976. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data." *Journal of the Royal Statistical Society* B38:290–95.

U.S. Department of Commerce. 1966. *Income Distribution in the United States*, by Herman P. Miller, a 1960 Census Monograph. Washington, DC: U.S. Government Printing Office.