"REFERENDUM" CONTINGENT VALUATION ESTIMATES: SENSITIVITY TO THE ASSIGNMENT OF OFFERED VALUES

by

T.A. Cameron
Department of Economics, UCLA

and

D.D. Huppert
National Marine Fisheries Service, La Jolla, CA

September 1988

Working Paper #519
Department of Economics
University of California
Los Angeles, CA 90024-1477

"REFERENDUM" CONTINGENT VALUATION ESTIMATES: SENSITIVITY TO THE ASSIGNMENT OF OFFERED VALUES

by

T.A. Cameron*
Department of Economics, UCLA
405 Hilgard Avenue
Los Angeles, CA 90024-1477

and

D.D. Huppert
National Marine Fisheries Service
Southwest Fisheries Center
8604 La Jolla Shores Drive
La Jolla, CA 92308

Contingent valuation (CV) methods are becoming increasingly popular for assessing the value of non-market resources and public goods. Different formats have been suggested for the hypothetical CV questions. Several studies have now compared the value estimates resulting from alternative formats and have speculated upon the reasons for observed discrepancies. These reasons now include a whole taxonomy of possible biases. We take a closer look at one CV format--the referendum--and demonstrate that simply the "luck of the draw" in assigning the referendum thresholds on individual questionnaires can produce a surprisingly wide variety of value estimates. We control for the behavioral biases which confound other comparison studies by using one sample of "payment card" CV data and simulating 200 samples of consistent referendum responses. We conclude that when referendum questions produce different value estimates than other formats, elaborate explanations for the apparent discrepancies may not be necessary.

* We thank Michael Hanemann, John Loomis, Robert Mendelsohn, Jane Murdoch and Matthew Wright for helpful comments and suggestions. We have also benefited from the comments of anonymous referees who read earlier, shorter descriptions of this research.

1. INTRODUCTION

Contingent valuation methods (CVM) are being used with increasing frequency to assess the value of non-market resources and public goods. These hypothetical market exercises can provide valuable information about the characteristics of demand for a commodity which is not presently traded in a real market. (See Cummings *et al.* (1986) or Mitchell and Carson (1988) for detailed discussions.) As CVM grows in popularity, practitioners of the technique (and policy-makers who use its findings) require standards by which to judge the accuracy and robustness of the empirical results. This paper isolates and examines an important statistical issue in CVM: the substantial small-sample variation in the estimates produced by one popular type of CVM question.

Several formats have been suggested for posing contingent valuation questions to survey respondents. Early CV studies simply asked participants to state the dollar amount they would be willing to pay for access to a resource or public good. However, since respondents lack market experience with these goods, this type of question is sometimes very difficult to answer. Alternative elicitation methods were therefore sought.

Sequential or iterative bidding methods, or auction techniques have been explored, but some of these are difficult to implement for mail or telephone surveys. They can also be quite time-consuming. There is also some suspicion that the final values elicited by these methods can be influenced by the starting points chosen.

One increasingly popular format is the "referendum" approach. Referendum CV questions (sometimes called closed-ended, dichotomous choice, or take-it-or-leave-it CV questions) are often used for surveys conducted by telephone, in person, or by mail. The respondent is offered *just one* randomly

assigned amount and their yes/no response is recorded. The valuation information in these responses is diffuse, so a large number of responses is required to identify the approximate valuation. However, the method has the advantage of placing very little stress upon the respondent because the pricing scenario mimics the take-it-or-leave-it market decisions made daily by most of us. Consequently, it is sometimes presumed that a relatively higher completed response rate can be expected.

Another popular format is the "payment card." With a payment card, the respondent is asked simply to peruse a range of values and to circle the highest amount they would be willing to pay.[1] From this, it can be inferred that the respondent's true "point" valuation lies somewhere in the interval between the circled value and the next highest option. Payment cards also conserve respondent effort because even a fairly detailed set of value "thresholds" can be visually scanned very quickly. Payment cards were originally designed to minimize the problem of "starting point" bias inherent in iterative bidding formats (see Mitchell and Carson, 1988).

A substantial literature has now evolved around the comparison of empirical value estimates produced by CVM techniques, as opposed to those derived from other standard methods (including "travel cost" methods and "hedonic" housing market studies). Examples include Brookshire, Thayer, Schulze and D'Arge (1982), Smith, Desvousges, and Fisher (1986) and Dickie, Fisher, and Gerking (1987). Several other studies, both published and unpublished, are reviewed in detail by Cummings, Brookshire, and Schulze (1986). The CVM value estimates are sometimes close to those produced by other methods, but sometimes they are quite different. In response to occasional large disparities, researchers have speculated upon possible reasons for the observed differences. A taxonomy of potential biases has

grown up: hypothetical bias, vehicle bias, starting point bias, information bias, nonresponse bias, sample selection bias, and strategic bias, to name a few. (See Cummings, *et al.*, 1986, Edwards and Anderson, 1987.)

A subset of these comparison studies has specifically compared the point estimates of resource values resulting from *referendum* questions with those resulting from payment card and open-ended ("name the amount") CV questions, from simulated markets, and from indirect methods such as the travel cost technique or the hedonic method. (For example, see Bishop, Heberlein, and Kealy, 1983, Sellar, Stoll, and Chavas, 1985, Boyle and Bishop, 1988, and Kealy, Dovidio, and Rockel, 1988. A few referendum studies also appear among the array of early comparison studies discussed in Cummings *et al.*, 1986.) Our research allows a reinterpretation of these comparison studies.

In contrast to these earlier studies, we specifically *do not* undertake a referendum study along with one or more alternative valuation methods and then compare the results. Using a common sample of respondents is certainly important because sample heterogeneity is eliminated. However, this usual strategy does not allow one to discriminate between differences in estimated value that result from systematic biases (such as those listed above), and purely statistical artifacts. We go further in this study by eliminating the possible confounding effects of the usual taxonomy of behavioral biases. This is accomplished by using *one* sample of real "payment card" response data in conjunction with 200 samples of simulated (but fully consistent) referendum responses.

We show that when the results from *just one* referendum question on a survey are compared with other valuation estimates, it should not be at all surprising that the valuation point estimates are substantially different. One should not infer *systematic* bias when comparing one set of referendum

results with other schemes. In fact, even if the underlying "true" values elicited by the two methods are *completely* identical, widely different point estimates can still result from something as seemingly innocuous as changing the arbitrary assignment of thresholds on referendum questionnaires.

Section II sketches an appropriate statistical procedure for the maximum likelihood (ML) estimation of "referendum" models. Section III outlines an ML procedure for a "payment card" regression model. Section IV reviews the dataset and the specification of the valuation (WTP) function. In Section V, we describe the simulation exercises which make the main point of this paper. Section VI examines the implications of our results for the findings of other researchers. Section VII offers some caveats and conclusions.

2. MAXIMUM LIKELIHOOD ESTIMATION WITH "REFERENDUM" DATA

We do not wish to prejudice the empirical results in this paper by using inappropriate estimation methods. Instead, we are careful to adopt consistent assumptions about the parametric form of the true population distribution of resource values across our two different estimation procedures.

For our referendum data, we use full information maximum likelihood estimation methods which take full advantage of all of the information available in referendum responses. (See Cameron and James, 1987a,b, Cameron, 1988a). Many earlier researchers have used conventional dichotomous choice logit models with referendum data, in a procedure which involves fitting a logistic "dose-response" curve and integrating the area bounded by the curve (see Bishop and Heberlein, 1979, and a restatement of their approach in Hanemann, 1984). The estimation method used in this paper is conceptually different, being a censored normal regression model.

We assume that the respondent's true valuation is Y_i, and that $\log Y_i = x_i'\beta + u_i$, where u_i is normally distributed with mean 0 and variance σ. Under

a willingness to pay (WTP) scenario, the respondent is offered a single threshold value t_i; if he is willing to pay this amount, we record $I_i = 1$ (if not, $I_i = 0$). Then we can presume that:

$$(1) \quad \Pr(I_i = 1) = \Pr(\log Y_i > \log t_i) = \Pr(u_i > \log t_i - x_i'\beta)$$
$$= \Pr(u_i/\sigma > (\log t_i - x_i'\beta)/\sigma)$$
$$= 1 - \Phi((\log t_i - x_i'\beta)/\sigma).$$

The log-likelihood function is then:

$$(2) \quad \log L = \sum_{i=1}^{n} \{ I_i \log [1 - \Phi((\log t_i)/\sigma - x_i'\beta/\sigma)] + (1 - I_i) \log [\Phi((\log t_i)/\sigma - x_i'\beta/\sigma)]\}.$$

The presence of $\log t_i$ allows σ to be identified so that the underlying valuation function, $x_i'\beta$, can be recovered. (Note that if $\log t_i = 0$ for all i, we have the conventional maximum likelihood probit model.) A full description of the elements of the gradient and the Hessian has been relegated to an Appendix.

Once the optimal values of β and σ have been attained, it is a simple matter to reconstruct fitted values of the transformed variable log Y. The conditional mean of log Y for any given vector of x variables will be $x_i'\beta$. However, if we wish to retransform back to the estimated conditional distribution for the variable Y itself, the mean will be given by $\exp(x_i'\beta)\exp(\sigma^2/2)$, where σ is an unbiased estimate of the true underlying population error variance.

3. MAXIMUM LIKELIHOOD ESTIMATION WITH "PAYMENT CARD" DATA

With payment card *interval* data, researchers sometimes assign the midpoint of the relevant interval as a proxy for the mean of the variable over that interval and employ OLS regression using these midpoints as the dependent

variable. We have argued elsewhere (Cameron and Huppert, 1988a, and Cameron, 1987) that it is preferable to employ more-efficient maximum likelihood (ML) estimation methods for regression models where the dependent variable is only measured on intervals of a continuous scale.

As in our referendum models, we again assume a *lognormal* conditional distribution for valuations. If the respondent's true valuation, Y_i, is known to lie within the interval (t_{li}, t_{ui}), then log Y_i will lie between log t_{li} and log t_{ui}. If log $Y_i = x_i'\beta + u_i$ and u_i is distributed normally with mean 0 and standard deviation σ, then we can standardize the range of values log Y_i occupies and state that:

$$
\begin{aligned}
(3) \qquad & \Pr(Y_i \subseteq (t_{li}, t_{ui})) \\
& = \Pr\left((\log t_{li} - x_i'\beta)/\sigma < z_i < (\log t_{ui} - x_i'\beta)/\sigma \right) \\
& = \Phi[(\log t_{ui} - x_i'\beta)/\sigma] - \Phi[(\log t_{li} - x_i'\beta)/\sigma],
\end{aligned}
$$

since z_i is the standard normal random variable and Φ is the cumulative standard normal density function. If we let z_{li} and z_{ui} signify the lower and upper limits, the corresponding log-likelihood function for a sample of n independent observations is:

$$
(4) \qquad \log L(\beta,\sigma \mid t_{li}, t_{ui}, x_i) = \sum_{i=1}^{n} \log \left[\Phi(z_{ui}) - \Phi(z_{li}) \right].
$$

Cameron and Huppert (1988a) provides the formulas for the gradients and the Hessian matrix associated with this likelihood function.

4. THE DATA

We employ a subset of the sample described in Cameron and Huppert (1988a,b) and in Thomson and Huppert (1987). These data are drawn from the NOAA National Marine Fisheries Service "Bay Area Sportfish Economic Survey" (BASES) of California's San Francisco Bay Area. The crucial valuation

question was worded as follows: "What is the MOST you would be willing to pay each year to support hatcheries and habitat restoration that would result in a doubling of current salmon and striped bass catch rates in the San Francisco Bay and ocean area if without these efforts your expected catch in this area would remain at current levels? (*Circle the amount*)." The listed values were $0, $5, $10, $15, $20, $25, $50, $75, $100, $150, $200, $250, $300, $350, $400, $450, $500, $550, $600, and "$750 or more."[2]

Other questions on the survey elicited information on the respondents angling activity (including details of their most recent three fishing trips), their level of angling skill, boat ownership, expenditures on fishing gear, employment status, and household income. Only *some* of the variables constructed from this information consistently make statistically significant contributions to explaining the value placed on the enhancement efforts. Descriptive statistics for these influential variables are presented in Table 1.

Our valuation question does not elicit information on the *height* of the inverse demand curve itself. Instead, the estimated value, $\exp(x_i'\beta)\exp(\sigma^2/2)$, is a measure of the increase in total surplus due to whatever *vertical shift* in the demand function is to be expected from a doubling of current overall salmon and striped bass catch rates.[3]

In some applications, the quality of the data set warrants the specification of a fully utility-theoretic valuation function. Cameron (1988b) is one example. The simple log-linear form employed here can nevertheless be interpreted as an approximation to *some* utility-theoretic form.[4] (See Huppert, 1988.) The ML estimates for the payment card valuation interval responses appear in Table 2.

Table 1. Descriptive Statistics (n = 342)

Variable	Description	Weighted[a] Mean (std. dev.)
MIDPT	midpoint of interval respondent selects on the payment card	57.98 (132.96)
log(MIDPT)	log (midpoint of interval on payment card)	3.115 (1.371)
TRIPS	# salmon and striped bass fishing trips in past 12 mo.	4.416 (5.449)
log(INC)	log($'000 hhld income using midpoint of reported interval)	3.602 (0.6544)
BTRIP	= 1 if all trips were striped bass fishing trips	0.3338
ADVCD	= 1 if advanced fishing ability	0.2813
[b]S-TARG	catch/trip of salmon on excl. salmon trips	0.7341 (1.046)

[a] weights were derived from cross-tabulations of home county by frequency of trip for both the original sample and the estimation sample.

[b] simple averages; values can be zero if no trips of the specified type were taken. Since data are retrospective over last three trips, we cannot simply use BTRIP to compute actual per-trip catch--some anglers will have reported mixed trip types.

Table 2. ML Interval Estimates versus Simulated Referendum Data
Implicit Dependent Variable: Log(WTP) (200 Samples, n = 342)

Variable	Payment Card ML Interval[a]	200 Referendum Samples "lower bound"		"upper bound"	
	point est. (asy. std. err.) (asy. t-ratio)	mean (std.dev.) ("t-ratio")	max min	mean (std.dev.) ("t-ratio")	max min
constant	2.530 (0.2896)[b] (8.736)	2.269 (0.7793)[c] (2.911)[d]	4.270 -0.0301	1.969 (0.7671) (2.567)	3.695 -1.203
TRIPS	0.02862 (0.0100) (2.861)	0.03569 (0.02087) (1.711)	0.1047 -0.01856	0.03144 (0.01836) (1.712)	0.08845 -0.02771
log(INC)	0.2531 (0.07656) (3.306)	0.2652 (0.1950) (1.360)	0.9555 -0.2377	0.3246 (0.1939) (1.674)	1.107 -0.1064
BTRIP	-0.4495 (0.1238) (-3.631)	-0.5217 (0.3184) (-1.639)	0.1556 -1.890	-0.4763 (0.2734) (-1.742)	0.3502 -1.356
ADVCD	0.2781 (0.1205) (2.308)	0.4211 (0.2978) (1.414)	1.379 -0.4273	0.4791 (0.2602) (1.842)	1.258 -0.2869
S-TARG	-0.1497 (0.05569) (-2.688)	-0.2008 (0.1394) (-1.441)	0.08853 -0.5815	-0.2454 (0.1412) (-1.737)	0.05741 -0.6658
σ	0.8724 (0.03752) (23.25)	1.195 (0.2666) (4.481)	2.317 0.7160	1.231 (0.2396) (5.135)	2.136 0.7706
max logL[e]	-655.62	-150.70 (14.88)		-160.90 (17.17)	

[a] see similar results in Cameron and Huppert (1988a).

[b] asymptotic standard errors and asymptotic t-test statistics (GQOPT output, GRADX routine).

[c] standard deviation of point estimates across 200 simulations.

[d] analogy to t-value computed using standard deviation of simulated-response point estimates across 200 samples.

[e] maximized values of the log-likelihood functions are not comparable, since the models are non-nested. Conceptually, it is easier to predict residence of an observation in the wider "intervals" defined by the referendum format.

All of the models described in this paper are fundamentally non-linear in parameters, so we use the package of FORTRAN subroutines called GQOPT (Goldfeld and Quandt) to obtain our parameter estimates. Rough optimization for all estimates reported in this paper was achieved by the DFP procedure. Fine-tuning of the estimates (to a convergence tolerance of 10^{-10}) was accomplished by GRADX, a quadratic hill-climbing algorithm.

5. THE RANGE OF VALUES FROM REFERENDUM DATA

Despite the fact that the valuation data collected in this survey were *actually* collected by means of payment card, we can use the sample information to simulate the value estimates which could have been obtained if the *identical* underlying value responses had been elicited by a referendum question. By conducting simulations, we abstract from the variety of behavioral distortions which might have been introduced if a completely separate referendum question actually *had* been asked on the survey. This is not to say that the values implied by the payment card are necessarily the "true" values, although we will treat them as such.

It should be appreciated that this is not a conventional Monte Carlo simulation of CV responses. In a very different study, Monte Carlo procedures might start with a "known" population valuation function. One could then append error terms from a known distribution to generate a large number of "random draws" from some known population of *anglers*. The different results produced by *one* arbitrary "payment card censoring" of the true data--as opposed to *one* arbitrary "referendum censoring" of the true data--could then be explored.[5] But what payment card would you choose? And what set of referendum thresholds? The first question (variability in value estimates as a function of payment card design) is taken up in Cameron and Huppert (1988b). The present study addresses the second problem: the exclusive effects of

different referendum threshold assignments. Randomness of these thresholds is fundamentally different than the usual source of variability in Monte Carlo studies, existing at a different level than ordinary sampling error. In this study, we take not only the "population" valuation function as given; we take the observed disturbances as "true." We then proceed to examine further randomness introduced into the process by the survey designer who enters threshold values on each questionnaire.

In this section, we begin with the actual payment card responses given by the survey participants and used in the valuation function estimates of the previous section. We use these responses to construct the simple binary "yes/no" responses we would have *expected* to get if the valuation question had been posed in the form of a *single* randomly assigned threshold value.[6] In reality, the form of the valuation question almost certainly has a systematic effect (of unknown magnitude) on the implicit true point value that the respondent chooses to convey. But this exercise controls for these additional distortions and considers only the statistical effects of the different levels of censoring of the valuation information. As we shall see, these distortions can be substantial. They may even overwhelm the array of behavioral biases discussed in the literature.

Given the format of the payment card actually used, there are twenty different single "thresholds" we might have assigned (counterfactually) for each respondent. There are a very large number of complete assignments that could have been made to the full set of respondents, so simulation techniques are required to generate descriptive statistics for the range of possible value estimates that might have been achieved under the hypothetical referendum format. More thresholds are typically offered at the *lower* end of the value spectrum, to increase the resolution in the range where it is

suspected from preliminary tests that a large proportion of the valuations lie. Therefore, we opt to generate threshold values from frequency distributions similar to the *observed* frequencies for the valuations implied by our respondents. There is no standard procedure commonly used by other researchers.

Even if we rely on the observed distribution of intervals selected by the respondents, there are still *two* alternative assumptions regarding the underlying distribution from which to draw the single thresholds to be offered to each respondent in our simulations of the referendum questions. The first distribution mimics the distribution of observed *lower* bounds on the chosen intervals. In this case, the threshold assignments (for each respondent in each simulation sample) were made as follows. First we sorted all the observations on the basis of the size of the lower bound of the interval selected on the payment card. For these sorted observations, we recorded the range of observation numbers for which each lower bound occurred. This gives us the absolute frequency of occurrence of each lower bound in our sample of 342 respondents.

We then generated 342 random integers (one for each respondent) on the range of 1 to 342. These random integers were used to randomly assign to each respondent a referendum threshold for each simulation sample of size 342. An example may help to illustrate this procedure. The lower bound of $15 was indicated on the actual payment cards (in the *sorted* sample) for observations numbered 105 through 153. If the random integer (on the range of 1 to 342) generated for any given respondent happened to take on a value in the range of 105 through 153, we assigned a referendum threshold (t_i) of $15 for that person.

The next step is to figure out what this person's response would have been had they simply been asked if they would be willing to pay this amount. To generate the respondent's probable "yes/no" response, we look at the interval they *actually* chose on the payment card. If the lower bound of the interval that the individual *actually* chose was $25 (for example), we would infer that this person would have responded "yes" ($I_i = 1$) to the question of whether they would be willing to pay $15.

The other possible assumption about the distribution of the randomly offered thresholds mimics the observed distribution of *upper* bounds on the chosen payment card intervals. The assignment of thresholds "drawn from" this distribution was analogous to the procedure for the first distributional assumption, except that the whole distribution of assigned single thresholds will be shifted up by one level.[7]

Two hundred simulated samples were generated in each case. Starting values for the maximum likelihood optimization algorithm were taken from the ML interval estimates for the true payment card data.

In conventional *single*-sample estimation, asymptotic standard errors for parameter estimates are used as a proxy for the degree of dispersion that we might expect in repeated sampling from the same population. In simulation experiments, we in fact generate a large number of random samples. This makes it possible actually to *calculate* the dispersion in the point estimates across different samples. The payment card asymptotic standard errors in the first column of Table 2 are not directly comparable with the calculated standard errors (across 200 experiments) reported in the second and third columns of the table. However, their magnitudes give us a good idea about the extent of the loss in statistical efficiency when we move from a payment card valuation question to a referendum valuation question, controlling for any other sources

of distortion due to the change in formats. The standard deviations of the point estimates across the 200 simulations are between two and two-and-a-half times the asymptotic standard errors of the payment card point estimates.

Dispersion is one concern; systematic bias is another. For this data set, the referendum slope estimates tend on average to overestimate the "true" ML interval slope estimates. Possibly, this could be an artifact of the algorithm for the assignment of thresholds. This apparent bias in point estimates under the referendum format may not be a robust result, although the two threshold assignment schemes we have considered are probably the most logical arbitrary choices.

The important lesson to be drawn from these simulation exercises is that, across replications, the degree of variation in some of the referendum parameter point estimates is very high. In Table 2, we provide "artificial t-statistics" loosely constructed from the mean point estimate divided by the standard deviation (across the 200 simulated sample replications). At the 5% level, these "t-ratios" fail to reject the zero hypothesis for every slope value in the two models. The maximum and minimum coefficient estimates found over the 200 replications are also quite telling. In all cases, the maximum and minimum parameter point estimates have different signs, even for the intercepts. This confirms that a *very* wide range of point estimates can result from referendum-style data collection, even in the complete absence of higher-level behavioral biases.

It is also informative to consider the range of possible implications (i.e. value derivatives with respect to explanatory variables, and fitted marginal valuations) that might be drawn across the different sets of randomly assigned thresholds in the simulation experiments. Table 3 provides the average derivatives of valuation with respect to each explanatory variable.

It also gives the marginal mean fitted willingness to pay--the quantity emphasized in most comparison studies.

The first column of Table 3 gives the single sample results for the payment card case; the second and third columns give average results across 200 artificial samples for the referendum case (using chosen "lower" and "upper" intervals bounds, respectively, as the distribution from which to draw the threshold "offers"). The top portion of the table shows that these value derivatives have substantially different averages than the estimates produced from the payment card data. In many cases, they differ by a factor of two. The difference stems in part from the fact that these derivatives consist of the relevant parameter point estimates multiplied by the fitted individual valuations, which are typically larger for the artificial referendum samples than for the original payment card data set.

The bottom part of Table 3 holds the crucial message for other researchers who have compared referendum questions with other methods. Focusing on the "lower bound" experiments, the *mean* of the 200 fitted sample average marginal WTP estimates for the referendum samples is $70.66, as opposed to the value of $46.08 determined for the actual payment card results.[8] These point estimates are different enough, and suggest a systematic *bias* in the referendum value estimates (at least in comparison with the payment card results). But it is more startling to note the magnitude of the standard deviation of valuations produced by the referendum experiments: $55.04. In fact, the range of point estimates of value occurring in the 200 "lower bound" experiments is from $29.51 to $551.39. These are dramatically different values. And any one of these sets of results could just as easily have been attained in any single trial. The results for the 200 "upper bound"

Table 3. Comparison of Valuation Derivatives and Average Values for Payment Card[a] versus Referendum[b] Data

	Payment Card	200 Referendum Samples			
		"lower bound"		"upper bound"	
		mean (std.dev.) ("t-ratio")	max min	mean (std.dev.) ("t-ratio")	max min
Sample average ∂WTP/∂x:					
TRIPS	$ 1.319 (0.4830)	$ 2.946 (4.624)	$49.47 -1.32	$ 2.200 (2.433)	$19.94 -1.14
log(INC)	11.67 (4.270)	19.38 (30.19)	341.38 -18.98	20.98 (18.49)	113.80 -3.96
BTRIP	-20.72 (7.585)	-42.80 (64.40)	17.19 -605.20	-32.32 (30.22)	14.39 -193.92
ADVCD	12.82 (4.692)	34.22 (53.76)	595.27 -29.54	33.10 (32.34)	229.21 -13.29
S-TARG	-6.901 (2.526)	-17.14 (26.52)	6.06 -257.27	-17.32 (18.18)	2.32 -133.80
Sample average fitted WTP:					
	$ 46.08 (16.87)	$ 70.66 (55.04)	$ 551.39 29.51	$ 62.60 (31.96)	$ 236.21 32.32

a For the actual payment card data, these fitted derivatives are based on the one actual sample. The derivatives vary across observations because they consist of the estimated coefficient times the fitted valuation (which varies across observations), so we report the sample averages. The standard deviations reflect the variation across 342 observations of these derivatives.

b For the simulated samples, we compute for each sample the sample average derivatives, but for each explanatory variable, there are 200 of these sample averages, one for each artificial sample. The figures in the table are the averages, standard deviations, maxima, and minima, *across 200 samples*.

experiments are analogous. The overall value estimates range from $32.32 to $236.21, with a mean of $62.60 and a standard deviation of $31.96.

6. IMPLICATIONS FOR THE FINDINGS OF OTHER STUDIES

Other comparisons of payment card and referendum valuation studies have focused upon the possibility that respondents may divulge fundamentally different values depending upon what question format is used. Theoretically, this is completely possible. Discrepancies could stem from different strategic incentives, different interpretations of questions, from sample selectivity, or from any of the other sources of behavioral bias mentioned in the introduction. Researchers have frequently uncovered different empirical estimates of resource values when different elicitation methods are used. In response, they usually try to attribute these differences to one or more of these biases.

The experiments in this paper illustrate that even without any systematic behavioral biases, simply the arbitrary assignment of referendum thresholds (combined with the greater censoring of referendum value information) could easily lead to a wide range of different value estimates. Our comparisons have been limited to a payment card format, since that format has provided the more-refined value information we treat as "true." We find that referendum value estimates can easily be vastly larger, or vastly smaller, than payment card value estimates. Even very big differences in the resource values implied by the two methods *could* be merely an artifact of the small-sample variability of referendum estimates as a function of threshold assignments. In some cases, then, perhaps no more complex explanation is required.

The range of values we find across our 200 simulated referendum data sets is consistent with the sizes of the discrepancies observed between

referendum responses and alternative valuation question formats in existing single-sample studies. These differences could emerge even *without* any of the additional behavioral distortions. For example, Bishop, Heberlein, and Kealy (1983) examine travel cost, simulated market, and contingent market estimates of the value of a permit to hunt Canada geese. For a simulated market, they obtain a willingness-to-sell (WTS) estimate of $63; for an open-ended (name the amount) contingent valuation question, the corresponding estimate is $68; for a referendum question, the value is estimated to be $101. In the light of our results, it is entirely possible that the "true" value is about $65. The standard deviation of the estimates produced by all possible threshold assignments for the referendum question could conceivably be about 2/3 this size, and the value of $101 would be a (perfectly plausible) one standard deviation away from the overall mean of values which could result.

Bishop *et al.* go on to speculate upon the reasons for differences in their value estimates by alternative methods. They state that "...people responded more strongly to real than to hypothetical dollars, forcing CV [referendum] WTS ($101) to exceed substantially the comparable simulated market figure ($63). One has to wonder why." They then suggest that "differences in the impact of commitment," or uncertainty, may provide a partial answer, and relate their explanations to arguments in the literature concerning hypothetical bias. But they mention that "[r]andom error alone is not sufficient to establish a bias." The results of our study lead us to agree strongly with this last incidental point. Any conclusion of systematic bias drawn from comparison of a *single* referendum question asked of the same sized sample as other valuation questions is insupportable. Referendum value estimates have much higher variance; substantially larger samples are required

before these values can be considered statistically comparable to values produced by other methods.

Sellar, Stoll, and Chavas (1985) also provide resource value estimates by the referendum method and by other techniques over one common sample of respondents. Their mean values for boat ramp permits for three survey regions by the travel cost method are $32, $102, and $13. The corresponding referendum value estimates are $39, $35, and $14. (Considering that their sample sizes were 70, 74, and 15, respectively, it is rather surprising that the estimates for the first and third regions are so close.)

Our study shows that it is possible to get a standard deviation in referendum point estimates that is 2/3 as large as the mean of these values across different threshold assignments. It could be, for the second region in the Sellar, Stoll, and Chavas study, that the mean of all possible referendum estimates is *indeed* $102. The single-sample observed value estimate of $35 *could* therefore lie comfortably within about one standard deviation away from that mean. Of course, it is more likely in a random assignment of thresholds that the referendum estimates will lie nearer to the "true" mean of all possible referendum estimates. This might account for the closer correspondence between the travel cost estimates and the referendum estimates in the other two regions.

Sellar *et al.* note the disparity between the travel cost and referendum estimates for their second region. Rather than recognizing that this could result simply from the statistical accident of threshold assignments, they offer another explanation. "The difference between the two surplus estimates for the [region] with the largest travel cost surplus estimate...suggests that respondents may be more reluctant to reveal their true willingness to pay through a direct method when their surplus value is high (e.g. $100), but at

lower levels this is not a problem." In contrast, our study suggests that the difference could be purely statistical; in a different assignment of thresholds, the discrepancy could easily vanish.

As another example, Boyle and Bishop (1988) report a mean value of willingness-to-pay to preserve scenic beauty along the lower Wisconsin River (using payment cards) of $29.36. The estimates are presumably derived using interval midpoints. Their estimate of mean value for their referendum questions is $18.88. Boyle and Bishop are careful to note in their conclusions that "the selection of a range and distribution of offers for dichotomous-choice-valuation questions can affect the resulting value estimates," but they do not address the possible extent of these effects.

Kealy, Dovidio, and Rockel (1988) test different CV methods of valuing a particular good, using referendum and open-ended formats. The final sample size for the study was 148. Only sample mean valuations were computed. Willingness-to-pay estimates for the chocolate bar used for the valuation questions range from $.79 to $.85 for the hypothetical referendum format. Values ranged from $.72 to $.80 for the open-ended format. Actual purchase decisions implied a referendum value of $.56. Systematic biases in the referendum estimates are consistent with the findings of our paper. The fact that Kealy *et al.* find referendum estimates repeatedly in the same range suggests that the variance in the range of possible referendum estimates (across different threshold assignments) may be modest in this case. Or, their similar results from repeated referendum surveys may just be lucky "draws" from the middle of the distribution. The next experiment might have produced a wildly different referendum estimate. One cannot tell. In any case, chocolate bars are a very familiar market consumption good, which may

allow more precise value estimates than can be achieved for purely non-market environmental resources.

On the whole, for researchers attempting to "validate" referendum CV estimates against other estimates obtained by indirect or simulation methods, getting a good match might be entirely due to the luck of the draw. If one's survey had assigned the threshold values differently, the referendum value estimates could have been very, very different. It is fortunate for the progress of research in non-market resource valuation that "accidentally" disparate estimates occasionally emerge from comparison studies. It has encouraged theorists to think very carefully about all the reasons why referendum contingent valuation methods might yield values which differ from those produced by real markets, by market experiments, or by indirect methods. In most cases, of course, all of the biases which have been argued to afflict referendum studies will probably be active to some degree. But likewise, a substantial portion of the difference may be simply a statistical accident.

7. CAVEATS AND CONCLUSIONS

This paper has emphasized the high *variance* to be expected in referendum CV estimates solely as a consequence of random threshold assignments. But we have also uncovered an apparent bias in the referendum estimates (relative to our payment card results). Clearly, this bias is also a statistical artifact, since all other behavioral sources of bias are explicitly controlled-for by our simulation approach. Perhaps this new bias is a general property of referendum estimates. However, it is possible that the apparent upward bias is merely a consequence of our choice of a distribution from which to draw "random" single thresholds to assign to the respondents. Nevertheless, our arbitrary distributions are as plausible as any that might be made in practice, based on pre-testing of a survey population.[9]

Fortunately, the *typical* extent of the bias evidenced by the artificial referendum data in this particular example is not terribly severe. On average, over a large number of repetitions of the referendum format with a single sample, the implied value will be reasonably close to that implied by payment card methods. (This is despite the fact that any *single* referendum sample could yield radically different value estimates.)

This study generates information that will be of use to the designers of questions for contingent valuation surveys and to those responsible for deciding upon sample sizes. Of course, research into additional sources of distortion induced by alternative question formats should certainly continue. But the smaller the sample, the greater will be the danger of obtaining misleading results from a single referendum question on a survey. Referendum survey discrepancies which have been attributed to a range of exotic biases, such as strategic behavior, vehicle bias, hypothetical bias, etc., may simply be statistical artifacts. In many cases, what appears to be a large positive bias from the referendum format could easily have been no bias, or even a negative "bias," had the threshold values been assigned differently.

The lesson to be learned from the experiments reported in this paper is that referendum surveys can easily produce substantially different estimates of value for a non-market resource than those implied by other valuation questions posed to the same respondents. However, this should *not* be treated as an indictment of referendum CV questions. On the contrary, it should be treated partly as evidence that a *much larger sample* is required to achieve a comparable level of accuracy with referendum questions. For practitioners, survey sponsors, and research evaluators, then, this study provides a good argument for insisting on the *largest* affordable sample if the referendum question format is to be used. If the budget for a large referendum sample is

not available, it is possible that more-robust results can be obtained if the payment card or iterative bidding format is used instead.

REFERENCES

Bishop, R.C., and Heberlein, T.A. (1979) "Measuring Values of Extra-Market Goods: Are Indirect Measures Biased?" *American Journal of Agricultural Economics*, 61, 926-930.

Bishop, R.C., Heberlein, T.A., and Kealy, M.J. (1983) "Contingent Valuation of Environmental Assets: Comparisons with a Simulated Market," *Natural Resources Journal*, 23, 619-633.

Boyle, K.J., and Bishop, R.C. (1988) "Welfare Measurements Using Contingent Valuation: A Comparison of Techniques." *American Journal of Agricultural Economics*, 70, 20-28.

Brookshire, D.S., Thayer, M.A., Schulze, W.D., and D'Arge, R.C. (1982) "Valuing Public Goods: A Comparison of Survey and Hedonic Approaches," *American Economic Review*, 72, 165-177.

Cameron, T.A. (1987) "The Impact of Grouping Coarseness in Alternative Grouped-Data Regression Models," *Journal of Econometrics (Annals)*, 35, 37-57.

Cameron, T.A. (1988a) "A New Paradigm for Valuing Non-market Goods Using Referendum Data: Maximum Likelihood Estimation by Censored Logistic Regression," *Journal of Environmental Economics and Management*, 15, 355-379.

Cameron, T.A. (1988b) "Empirical Discrete/Continuous Choice Modeling for Non-Market Resource Valuation. Department of Economics, University of California at Los Angeles, Discussion Paper # 503, September.

Cameron, T.A., and Huppert, D.D. (1988a) "OLS Versus ML Estimation of Non-market Resource Values with Payment Card Interval Data," *Journal of Environmental Economics and Management*, 15, 355-379.

Cameron, T.A., and Huppert, D.D. (1988b) "The Sensitivity of Contingent Market Valuation Estimates to the Design of a Payment Card," mimeo, Department of Economics, University of California, Los Angeles.

Cameron, T.A., and James, M.D. (1987a) *The Determinants of Value for a Recreational Fishing Day: Estimates from a Contingent Valuation Survey*, Canadian Technical Report of Fisheries and Aquatic Sciences No. 1503, Regional Planning and Economics Branch, Department of Fisheries and Oceans, Vancouver, British Columbia, Canada.

Cameron, T.A., and James, M.D. (1987b) "Efficient Estimation Methods for Use with 'Closed-Ended' Contingent Valuation Survey Data," *Review of Economics and Statistics*, 69, 269-276.

Cummings, R.G., Brookshire, D.S. and Schulze, W.D. (1986) *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*, Totowa, NJ: Rowman and Allanheld Publishers.

Dickie, M., Fisher, A., and Gerking, S. (1987) "Market Transactions and Hypothetical Demand Data: A Comparative Study," *Journal of the American Statistical Association*, 82, 69-75.

Edwards, S.F., and Anderson, G.D. (1987) "Overlooked Biases in Contingent Valuation Surveys: Some Considerations," *Land Economics*, 63, 168-178.

Goldfeld, S. and Quandt, R.E. (undated) *GQOPT*: A Package for Numerical Optimization of Functions, Department of Economics, Princeton University, Princeton, N.J. 08544.

Hanemann, W.M. (1984) "Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses," *American Journal of Agricultural Economics*, 66, 332-341.

Huppert, D.D. (1988) "An Examination of Nonresponse Bias and Divergence Among Value Concepts: An Application to Central California Anadromous Fish Runs," Draft report, National Marine Fisheries Service, Southwest Fisheries Center, La Jolla, CA 92037 (May 26 version).

Kealy, M.J., Dovidio, J.F., and Rockel, M.L. (1988) "Accuracy in Valuation is a Matter of Degree," *Land Economics*, 64, 158-171.

Mitchell, R.C. and Carson R.T. (1988) *Using Surveys to Value Public Goods: The Contingent Valuation Method*, Resources for the Future, in press.

Sellar, C., Stoll J.R., and Chavas J.-P. (1985) "Validation of Empirical Measures of Welfare Change: A Comparison of Nonmarket Techniques," *Land Economics*, 61, 156-175.

Smith, V.K., Desvousges, W.H., and Fisher, A. (1986) "A Comparison of Direct and Indirect Methods for Estimating Environmental Benefits," *American Journal of Agricultural Economics*, 68, 280-290.

Thomson, C.J. and Huppert D.D. (1987) *Results of the Bay Area Sportfish Economic Study (BASES)*, U.S. Department of Commerce, NOAA Technical Memorandum, National Marine Fisheries Service.

APPENDIX: DERIVATIVES FOR THE "REFERENDUM" MODEL

Using the notation established in the text, we first define the following simplifying abbreviations (z denotes the standard normal random variable in this appendix):

$$z_i = (t_i - x_i'\beta)/\sigma$$

$$\Phi_i = \Phi(z_i) \qquad \phi_i = \phi(z_i) \qquad \phi'_i = \phi'(z_i) = -z_i\,\phi(z_i)$$

$$R_i = x_{ir}x_{is}\,\phi'_i \qquad S_i = x_{ir}x_{is}\,\phi^2_i$$

$$T_i = x_{ir}z_i\,\phi'_i \qquad U_i = x_{ir}z_i\,\phi^2_i$$

$$V_i = z^2_i\,\phi'_i \qquad W_i = z^2_i\,\phi^2_i$$

The gradient vector for this model is then given by:

$$\partial \log L/\partial\beta_r = (1/\sigma)\sum\{\,[I_i - (1 - \Phi_i)]\,x_{ir}\,\phi_i\,/\,[\Phi_i\,(1 - \Phi_i)]\} \qquad r = 1,\ldots,p$$

$$\partial \log L/\partial\sigma = (1/\sigma)\sum\{\,[I_i - (1 - \Phi_i)]\,z_i\,\phi_i\,/\,[\Phi_i\,(1 - \Phi_i)]\}$$

The elements of the Hessian matrix can be simplified if we define the function:

$$G(P,Q) = \sum\left[\frac{I_i(P_i[\Phi_i - 1] - Q_i)}{[\Phi_i - 1]^2} + \frac{(1 - I_i)(P_i\Phi_i - Q_i)}{\Phi^2_i}\right]$$

Then:

$$\partial^2 \log L/\partial\beta_r\partial\beta_s = (1/\sigma)\,G(R,S) \qquad r,s = 1,\ldots,p$$

$$\partial^2 \log L/\partial\beta_r\partial\sigma = (-1/\sigma)\,\partial \log L/\partial\beta_r + (1/\sigma^2)\,G(T,U) \qquad r = 1,\ldots,p$$

$$\partial^2 \log L/\partial\sigma^2 = (-1/\sigma)\,\partial \log L/\partial\sigma + (1/\sigma^2)\,G(V,W)$$

Use of these analytic derivatives instead of numerical approximations can reduce computational costs.

NOTES

[1] Willingness to accept can be elicited similarly.

[2] Respondents who circled $0 were also asked: "Did you circle $0 because you feel this change has no value to you?" If a respondent replies "no" to this question, we assume that this answer indicates that they value the change by more than $0 but by less than $5. While these could be "protest" bids, we do not pursue the matter here.

[3] This is expressly not a doubling of the number of fishing days, the "quantity" of the good for which demand is being modeled, but a doubling of one of the factors that is argued to "shift" this demand curve.

[4] The point made in this paper concerns the *format* of the valuation question and the *size* of the sample. Our basic findings are independent of the choice of functional form.

[5] This line of inquiry was suggested by one referee. These set of experiments would certainly be informative for practitioners, although they would address issues quite different from the one examined in this paper. For the problem of interval data for a dependent variable--the sort of data deficiency produced by payment card value elicitation--a Monte Carlo assessment of the biases produced by reliance on OLS estimation using interval midpoints is contained in Cameron (1987). The implications of this basic Monte Carlo work for actual payment card data are explored in Cameron and Huppert (1988a).

[6] Each referendum threshold divides the range of possible valuations into two "intervals." However, the referendum approach differs from the payment card method in that these thresholds are varied across respondents. With the simplest payment card strategy, everyone receives the same set of thresholds. It is the fact that the referendum thresholds vary which allows us to identify the location and scale of the conditional distribution of valuations.

[7] We initially assigned each of the twenty thresholds according to the outcomes from a *uniform* random integer generator (on the range of 1 to 20). This led to estimation problems, though, because by far the largest proportion

of respondents implied valuations among the lower intervals. This resulted in a very large proportion of "no" responses in the simulated samples because the occurrence of high-valued thresholds was drastically disproportionate relative to the actual valuations in the sample. Some of the simulations failed to converge because of this.

[8] Of course, this degree of precision overstates the accuracy of contingent valuation methods, but we elect to report the point estimates as money values.

[9] In practice, it seems important to conduct preliminary small-scale trial surveys to determine whether the intended range of offered thresholds "cover" the range of values existing in the population. Ill-conditioned data (such as a sample wherein a vastly disproportionate number of "yes" and "no" referendum responses are collected) can be expected to yield questionable results.