A Meta-Analysis of Single-Bound Contingent Valuation: Willingness to Pay Estimates, Determinants of Reliability and Split-Sample Hypothesis Tests

John C. Whitehead, Appalachian State University

Tim Haab, The Ohio State University

Lynne Lewis, Colorado State University

Leslie Richardson, National Park Service

Pete Schuhmann, University of North Carolina Wilmington

September 2, 2024

Abstract. The single-bound valuation question, often applied in the context of hypothetical referenda, is considered to be incentive compatible when surveys are consequential with a coercive payment vehicle and has become the preferred question format for contingent valuation studies. Yet, it provides a minimal amount of information with which to estimate willingness-to-pay (WTP) and its determinants. This problem can lead to unreliable WTP point estimates and wide confidence intervals. In this paper we investigate the problems associated with single-bound contingent valuation with a meta-analysis data set initially constructed by Lewis, Richardson and Whitehead (2024) for nonparametric WTP estimates. We extend this analysis to parametric estimates of WTP. With these data we find that parametric WTP estimates can differ substantially from nonparametric Turnbull WTP estimates. Smoothed Turnbull data produces confidence intervals that are significantly tighter than those from parametric WTP estimates using the Delta Method and Krinsky-Robb approaches. In a meta-regression, we estimate the magnitude of the inefficiency from single-bound data is increasing in the percentage of non-monotonicities and the flatness of the tail of the distribution. We demonstrate the problems created by these differences with directional split-sample t-tests from the meta-data. Tests with the Turnbull WTP estimates are more likely to find statistically significant differences relative to tests of differences in means with symmetric confidence intervals.

Introduction

The contingent valuation method (CVM) is a stated preference approach to the valuation of public goods for benefit-cost and other types of policy analyses (Carson 2012). CVM began with attempts to directly elicit consumer surplus with open-ended statements of value (Brown and Hammock 1973). However, following the introduction of the dichotomous choice response format by Bishop and Heberlein (1979), single-bound contingent valuation questions have remained the preferred question format. Single-bound questions present a policy with a single cost and yes/no answer categories to survey respondents in the context of a purchase of a product, a quasi-public good (e.g., a recreation trip) or support of a policy. In the case of public goods, the question format evolved to a for/against vote in the context of a policy referendum.

A number of influential publications have led to dominance of the closed-ended/dichotomous choice/single bound question format in the CVM literature. Hanemann (1984) developed the indirect utility theory to support the use of single-bound data. Cameron and James (1987) and Cameron (1988) developed the expenditure difference approach (now called "estimation in willingness to pay space" in the discrete choice experiment literature). Mitchell and Carson (1989) describe the advantages of framing the dichotomous choice question as a referendum. McConnell (1990) compared the theoretical properties of the indirect utility and expenditure difference approaches and Loomis and Park (1992) compared them empirically. The NOAA Panel (Arrow et al. 1993) endorse the referendum format for national resource damage assessment.[1] Carson and Groves (2007) provide a theoretical base to claim that a consequential

[1] Page 24 of the mimeo: "The above considerations suggest that a CV study based on the referendum scenario can produce more reliably conservative estimates of willingness to pay, and hence of compensation required in the aftermath of environmental impairment, provided that a concerted effort is made to motivate the respondents to

referendum question with a coercive payment vehicle is incentive compatible. Carson, Groves and List (2017) conduct an experimental test of the incentive compatibility of the single bound question with consequentiality to further bolster those claims. Finally, in a paper on best practice recommendations for stated preference studies to support decision making, Johnston et al. (2017) recommend the use of the single bound question based on the established incentive properties and empirical evidence regarding the validity of responses derived using this format. This body of research has led to a consensus that the single-bound CVM question is the "gold standard" for value elicitation. And yet, the data that results from surveys that employ single-bound questions are often problematic.

Econometric approaches to estimation of willingness to pay (WTP) with single-bound data has generated a large literature. Haab and McConnell (2003) spend approximately one-third of their econometrics of non-market valuation book (which includes travel cost methods and hedonic pricing) on dichotomous choice contingent valuation.[2] Haab and McConnell emphasize that single-bound valuation questions provide only a minimal amount of information with which to estimate WTP and its determinants. The researcher only learns if the respondent values the policy above or below the randomly assigned cost amount. Problems arising from single-bound data include negative WTP estimates, non-monotonicities and fat/flat tails. Each of these empirical issues will decrease the accuracy and statistical efficiency of WTP estimates.

Negative WTP estimates will result when the estimated probability of a yes/for response is less than 50% at the lowest cost amount and probit or logit models are used for estimation

take the study seriously, to inform them about the context and special circumstances of the spill or other accident, and to minimize any bias toward high or low answers originating from social pressure within the interview."

[2] See also Hanemann and Kanninen (2001).

(Hanemann 1984, Haab and McConnell 1997). Hanemann (1989) provides a formula for estimating WTP that eliminates this negative portion. Another common response to this problem has been to estimate the probability of a yes response with a log cost functional form. This model produces an estimate of median WTP but the mean WTP is often undefined (Haab and McConnell 2002). The Turnbull (Haab and McConnell 1997) assumes a lower bound on WTP at zero. While avoiding negative estimates of expected WTP by assumption, for reasons we will see below, the Turnbull in fact, cannot provide an estimate of expected WTP without imposing additional assumptions about the distribution of WTP between bids, and in the upper tail, above the highest bid. Kriström (Kriström 1990) and linear probability models provide estimates of mean WTP by imposing the non-negativity assumption as well as distributional and upper tail assumptions.

Non-monotonicity results when the probability of voting for the policy rises when the cost amount rises in pairwise cost comparisons. Haab and McConnell (2003) call this the "difficult data" situation. This violation of rational choice theory may simply be a result of sampling error due to small samples at each of the cost amounts. The Turnbull and Kriström nonparametric approaches handle this problem by pooling cost amounts and yes/for responses until the probability of the yes/for function is monotonically decreasing, or flat, as cost amounts increase. The logit, probit and linear probability models smooth the data by estimating a constant slope over the entire range of cost amounts. Beyond the problem of a lack of theoretical validity, non-monotonicities will lead to increasing standard errors of WTP.

The fat tails problem exists when the probability of a yes response is relatively high, say 20% or more, at the highest cost amount (Parsons and Myers 2016, Lewis et al. 2024). Fat tails leave the researcher uncertain about a potentially large portion of the WTP distribution. The Turnbull estimator deals with fat tails by ignoring the upper tail all together, resulting in the Turnbull failing to provide a point estimate of expected WTP and instead only providing a statistical lower bound on expected WTP. These lower bound WTP estimates may be appropriate for natural resource damage assessment and sensitivity analysis in benefit-cost analysis but are less appropriate as estimates of the central tendency of WTP (Lewis, Richardson and Whitehead 2024). The Kriström and linear probability models deal with fat tails by trying to identify the cost amount that leads to a zero probability of support. The linear probability model estimates this by forecasting beyond the range of costs. The use of probit and logit models can lead to WTP estimates that are greater than the highest cost when the data suffer from fat tails. Related, the flat tail problem exists when the when the probability of a yes response is relatively flat at two or more cost amounts (Lewis, Richardson and Whitehead 2024). Flat tails lead to less precise WTP estimates.

Our contribution is to investigate these problems with single-bound contingent valuation questions with a meta-analysis data set initially constructed by Lewis, Richardson and Whitehead (2024) for a comparison of nonparametric WTP estimates. We extend the analysis of these data to parametric estimates of WTP. We find that the parametric estimates of expected WTP differ substantially from the lower bounds on expected WTP produced by the Turnbull.[3] We then calculate standard errors and show that the inefficiency of single-bound data is increasing in the percentage of non-monotonicities over the range of cost amounts, as well as with the fatness and flatness of the tail of the distribution. Smoothed Turnbull data tends to produce confidence intervals that are significantly tighter than parametric WTP estimates from

[3] This result is not new (e.g., Bengochea-Morancho, Fuertes-Eugenio, and Saz-Salazar 2005).

the Delta Method and Krinsky-Robb confidence intervals. We demonstrate the problems created by these differences with 52 directional split-sample tests from 16 studies in the meta-data. The Turnbull WTP estimates are more likely to lead to failure of rejection of null hypotheses of no effect relative to t-tests of differences in means with the Delta Method standard errors and Krinsky-Robb confidence intervals. While somewhat interesting, this is to be expected as the Turnbull is providing a point estimate of expected WTP, but rather providing an estimate of the statistical lower bound on expected WTP, a point that is often confused in the literature.

In the next section we first review the theory and estimation of single-bound contingent valuation methods. We illustrate the various approaches to WTP estimation and show that WTP estimates are reliable over various estimation approaches with well-behaved textbook data. We then proceed as described above and, after discussing issues with double-bound contingent valuation questions, conclude with a possible direction forward.

Single-Bound Contingent Valuation

Suppose a consumer has a willingness to pay for a change in the quality or quantity of a public or quasi-public good, $WTP(\Delta q)$. A single-bound contingent valuation question would ask the respondent something like, “are you willing to pay $\$A$ for Δq?”, where $\$A$ is a randomly assigned cost amount. Since Carson and Groves (2007), the question is typically posed as a referendum vote, and if respondents consider the survey to be consequential with a coercive payment vehicle (e.g., a tax), the responses to the question are considered to be incentive compatible. Another type of question is posed for goods that are quasi-public, such as a recreation trip: “would you still take the trip if it cost an additional $\$A$?” (e.g., Cameron 1988). The theory and estimation methods are the same as in the referendum question and the questions are incentive compatible since there is no reason to strategize when a government policy is not involved. These questions can be extended to trips with a quality change (e.g., Neher et al. 2017). Note that Carson et al. (1996) find that estimates from this type of CVM question are convergent valid with estimates from the travel cost method, while Carson and Groves (2017) do not address this type of CVM question.

The consumer will answer yes/for to the single-bound valuation question if their willingness to pay is greater than or equal to the cost amount: $Pr(yes) = Pr(WTP \geq A)$. The Turnbull non-parametric estimator (Haab and McConnell 1997) produces a lower bound on mean WTP by assuming non-negative WTP: $WTP0 = \sum_j A_j \times \left[\Pr\left(yes_j\right) - \Pr\left(yes_{j+1}\right)\right]$, where $\Pr\left(yes_{j-1}\right) = 0$ at the lowest cost amount. Essentially, the Turnbull calculates the sample proportion of respondents falling between bid amounts and assigns every between bid proportion a WTP equal to lower bound on that bid range. This means that any ‘yes’ response to the highest bid is assumed to have a WTP no greater than the highest bid, and any ‘no’ response to the lowest bid is assigned a WTP of zero. The Kriström (1990) nonparametric estimator uses linear interpolation to estimate the choke price and assumes a zero percentage yes response at the interpolated choke price. Lewis, Richardson and Whitehead (2024) find that the Kriström WTP estimates are susceptible to the “fat tails” problem and propose a correction. In this paper, we focus our comparison on the Turnbull and parametric WTP estimates.

Hanemann (1984) provided theoretical justification for single-bound contingent valuation. Beginning with a linear utility function, Hanemann shows that the logit (and probit) model relies on the notion that willingness to pay is equal to the cost amount that makes respondents indifferent between voting for or against the policy (i.e., $Pr(yes) = 0.50$). If the logistic regression model includes only the cost amount as a determinant, $\Pr(yes) = 1/(1 + \exp(-(\alpha + \beta A))$, then the willingness to pay estimate is $WTP1 = -\alpha/\beta$. In a linear-in-cost model, the mean willingness to pay is equal to the median.

The logit model can also produce a section of negative WTP if the estimated logistic regression curve intersects the probability of a yes response axis below 100%. If the intersection is below 50% ($\alpha < 0$) then $WTP1 < 0$. Hanemann (1989) proposed a widely used correction that truncates the negative portion of the WTP distribution: $WTP2 = (-1/\beta) \times \ln(1 + \exp(\alpha))$, however, as Haab and McConnell (2002) note, the Hanemann procedure does not produce a statistically valid WTP estimate as the arbitrary correction results in a WTP distribution that does not integrate to one.

A logged cost amount model, $\Pr(yes) = 1/(1 + \exp(-(\alpha + ln\beta))$ has also been used to solve the "negative WTP" problem. The resulting WTP estimate is the median, $WTP3 = exp(-(\alpha/\beta))$, with an undefined mean WTP in the logit. A probit model with a logged cost amount can be used to find the mean WTP under certain conditions. Finally, willingness to pay can also be estimated from a linear probability model: $\Pr(yes) = \alpha + \beta A$, although, to our knowledge, this has not appeared in the peer-reviewed literature (Loomis 1988). Intuitively, WTP is the triangle under the regression line: $WTP4 = 0.5 \times \alpha \times (-\alpha/\beta)$. The area of this WTP triangle is the linear version of the nonparametric Kriström WTP estimate and closely approximates $WTP2$.

To illustrate the WTP estimation methods we consider some textbook CVM data. We construct a data set from question number 3 from Boardman et al.'s (2015) chapter on the contingent valuation method (p. 398). Students are told to "consider a project that would involve purchasing marginal farmland that would then be allowed to return to wetlands capable of supporting migrant birds. Researchers designed a survey to implement the dichotomous choice method. They reported the following data." In the data table there are ten costs that range from \$5 to \$50 and the percentage of those who are willing to pay each cost falls from 91% to 2%. Students are asked "What is the mean WTP for the sampled population?"

We create the data with 100 observations at each of the 10 cost amounts. The Turnbull $TWTP$ is estimated in MS Excel. The logit and OLS regression models of cost on the yes/no responses are $\Pr(yes) = 1/(1 + \exp\left(-(2.80 - 0.14 \times A)\right))$ and $\Pr(yes) = 0.95 - 0.021 \times A$, respectively. The willingness to pay estimates (with standard errors in parentheses) are \$18.50 (0.57), \$20 (0.64), \$21 (0.59), \$18 (0.61) and \$21 (0.55) from the Turnbull ($TWTP$), logit ($WTP1$-$WTP3$) and linear probability ($WTP4$) models, respectively. The willingness to pay estimates are not statistically different across valuation method. Unfortunately, as we will see, real world single-bound data are not so well-behaved.

Meta-Data

Parsons and Myers (2016) reviewed eight journals from 1990 to 2015 and found 86 articles that reported the percentage of yes responses at the highest cost amount. Forty-six of these articles provide the information necessary to reconstruct the data (Lewis et al. 2024). In addition to these studies, Lewis et al. (2024) searched the same set of journals for articles published through 2023 and found five additional articles that contain the necessary information to reconstruct the relevant data.

The data summary by study is presented in Table 1. The articles were published between 1990 and 2022 with all but three between 1995 and 2018. Sixty-one percent (31) of the studies are U.S. based with 5 studies based in Sweden, 3 in Spain, 2 in England and 1 each in Australia, Austria, China, Ireland, Kuwait, Mexico, the Philippines, Taiwan, Uruguay, and Vietnam. Twenty-two percent of the articles use a donation or voluntary contributions payment vehicle. There are five survey modes represented in the sample with the percentages adding up to more than one due to mixed modes being used in three studies. Forty-seven percent of the studies used a mail survey contact mode, 25% used an in-person contact mode, 14% are laboratory experimental modes (with student samples), 14% are telephone survey modes and 6% are online surveys. Seventy-one percent of the studies are valuing public goods. Fifty-three percent have one-time payment schedules. The average number of years in each payment schedule is 8 with a range of 1 (for one-time payments) to 30, where in perpetuity payment schedules are coded as 30.

Of these 51 articles, 21 have only one data set and the remainder have between 2 and 9 data sets. Twelve articles have 2 data sets, 10 articles have 3 data sets, 4 articles have 4 data sets, 2 articles have 6 data sets, 1 article has 8 data sets and another has 9 data sets. In total, there are 120 data sets available for analysis. In those articles that present multiple data sets the source could be an experimental treatment or samples of different populations. The mean sample size is 433 with a range of 47 to 4361 (Table 2). The average number of cost amounts presented to respondents is 7 with a range of 3 to 21. The mean of the sample size per cost amount is 71 with a range of 7 to 396. Twenty-two percent of the pairwise comparisons of yes responses to cost amounts exhibit non-monotonicities over the 120 data sets.

The cost amounts are left in the home country currency and not adjusted for inflation so the cost amounts themselves contain a limited amount of information. In order to make the bid amounts comparable across studies, for each individual study, we divide each bid amount by the maximum bid amount so that the standardized bid amounts can range from zero to one. The mean of the standardized minimum bid is 0.10. The two bids that form the slope for the tail of the distribution are the two highest bids inclusive of bids pooled for non-monotonicity. Forty-four percent of the data sets have pooled bids for one of the bid amounts used to calculate this slope. The mean of the standardized low bid in the slope (Sbid1) is 0.56 and the mean of the standardized high bid in the slope (Sbid2) is 0.88. The average percentage yes response at Sbid1 (Pctyes1) is 35% and the average percentage yes response at Sbid2 (Pctyes2) is 23%. The absolute value of the slope with the standardized bids is 0.48 with a range of 0.01 to 4.02.

WTP Estimates

We construct the Turnbull $WTP0$ estimates in MS Excel and the mean $WTP1$ and $WTP2$, median $WTP3$ and linear probability $WTP4$ estimates from logit and linear probability models for each of the 120 data sets (Table 3). One of the median $WTP3$ estimates approached infinity so it is dropped from the data summary. As expected, the Turnbull lower bound on mean WTP, $WTP0$, is 332, lower than all of the other WTP estimates except mean $WTP1$ for which 18% of the values are negative. The parametric median $WTP3$ estimate is 40% higher than the Turnbull lower bound estimate. The ratio of the truncated mean $WTP2$ to the median $WTP3$ estimate is 2.37 indicating significant variability across functional form. The truncated mean $WTP2$ and linear $WTP4$ estimates are, not surprisingly, very similar since both estimates lop-off the negative portion of the WTP distribution. We next delete the WTP estimates for which the

Mean $WTP1$ estimate is negative. Of the remaining 99 samples, the WTP estimates are much closer in magnitude.

For the 99 WTP estimates with positive $WTP1$ values we construct the ratios of the different WTP estimates to the others (Table 4). We find large differences. As expected, the mean of the $WTP2$ is larger than the Turnbull lower-bound $WTP0$. The ratio is 1.73 with a range of 1 to 12. The mean of the mean $WTP2$ to mean $WTP1$ ratio is 4.06 with a range of 1 to 210. The mean of the $WTP1$ to median ratio is 1.30 with a range of 0 to 4. The mean of the $WTP2$ to median $WTP3$ ratio is 2.51 with a range of 0 to 16. Twenty-three percent of the $WTP1$ to median $WTP3$ ratios are less than 1 and 5% of the $WTP2$ to median $WTP3$ ratios are less than 1. The ratio of the WTP1 to linear WTP estimates is 1.09 with a range of 0.69 to 1.33. This mean is not statistically different from 1 suggesting that the linear WTP estimate is a useful approximation and alternative to the mean $WTP2$. In regression models we are unable to explain the variation in these ratios with the percentage of pooled bids or sample size as independent variables.

Confidence Intervals

We estimate the standard errors of the Turnbull $WTP0$ estimates with the formula found in Haab and McConnell (p. 75, 2002). Standard errors of the parametric WTP estimates are calculated using the Delta Method, a first-order Taylor Series expansion from the variance-covariance matrix (Cameron 1991).

The t-statistics, $t = WTP/SE$, are significantly higher for the Turnbull $WTP0$ estimates relative to the parametric estimates (Table 5). This is due to the difference in methods used to construct standard errors (see the differences in the textbook data), as well as the facts that the Turnbull survival function is smoothed when non-monotonicities are encountered and because the Turnbull survival function does not have a fat/flat tail. Non-monotonicities and fat/flat tails will increase the standard errors of the slope coefficient in regression models. This coefficient is in the denominator of WTP estimates so the standard error of WTP estimates will increase as well.

We test for positive and statistically significant WTP estimates for each of the estimation methods. The significance level is 90% in a one-tailed test and the critical value is $t = 1.282$. All of the Turnbull $WTP0$ estimates are statistically significant. In contrast, 13% (n=13) of the non-negative mean WTP1 estimates are not statistically different from zero. Combined with the negative $WTP1$ estimates, 28% of the $WTP1$ estimates are not useable for policy analysis. Ten percent of the median $WTP3$ estimates are not statistically different from zero and 2.5% of the mean $WTP2$ and $WTP4$ estimates are not statistically different from zero.

The distribution of a ratio of parameters (such as WTP) is not necessarily symmetric. The asymmetry gets more severe when the parameter in the denominator is imprecisely estimated. Another approach to estimating confidence intervals that is common in the contingent valuation literature and captures this asymmetry is the Krinsky-Robb approach (Park, Loomis and Creel 1991). The Krinsky-Robb confidence interval is based on a simulation from the variance-covariance matrix of the estimated parameters and does not impose symmetry. Hole (2007) compares the Delta Method and Krinsky-Robb approaches and finds little difference for well-behaved (simulated and real) data. However, Hole (2007) points out that WTP must be normally distributed for the Delta Method confidence interval to be accurate.

We estimate the Krinsky-Robb confidence intervals in SAS software. We simulate one million WTP estimates and trim the lowest 2.5% and highest 2.5% values to estimate the 95% confidence interval. Krinsky-Robb confidence intervals are significantly wider than Delta Method confidence intervals (Table 6). Of n=98 estimates where mean *WTP1* is greater than 0, one of the ratios of the Krinsky-Robb confidence interval to the Delta Method confidence interval is less than 1 and one ratio is greater than 64. Trimming these 2 ratios, the mean ratio is 1.41 with a range of 1.01 to 5.51. Similarly trimming one ratio less than 1 and one ratio greater than 64, the mean of the ratio of Krinsky-Robb to Delta Method 95% confidence interval for $WTP2$ is 1.65 with a range of 1 to 6.32.

We test for positive and statistically significant WTP estimates for mean $WTP1$ and mean $WTP2$ by determining if the Krinsky-Robb confidence interval includes zero. Forty-four percent of the mean $WTP1$ Krinsky-Robb confidence intervals include zero. In contrast, only 18% of the Delta Method non-negative mean $WTP1$ confidence intervals include zero. Ten percent of the mean $WTP2$ Krinsky-Robb confidence intervals include zero. Only 3% of the Delta Method non-negative mean $WTP2$ confidence intervals include zero.

Data problems may also lead to asymmetries in the Krinsky-Robb confidence intervals. For those confidence intervals that do not include zero we measure asymmetry by $Asymmetry = (U95KR - meanWTP)/(meanWTP - L95KR)$, where $U95KR$ is the upper 95% Krinsky-Robb bound and $L95\ KR$ is the lower 95% Krinsky-Robb bound. The Krinsky-Robb asymmetry ratio for mean $WTP1$ is 1.50 with a range of 0.49 to 6.94 (Table 7). The Krinsky-Robb asymmetry ratio for mean $WTP2$ is 2.66 with a range of 1.19 to 8.41.

Meta-regressions

In order to test our contention that non-monotonicities and fat/flat tails contribute to statistical inefficiencies, we estimate a linear regression model with the Delta Method t-statistic as the dependent variable (Table 8). The independent variables are the percentage of the number of pooled bids, the height and slope of the tail, and sample size. The standard errors are clustered at the study level.[4] Each of the regression models are statistically significant at the $p < 0.01$ level and the R^2 values suggest that between 24% and 61% of the variation in the t-statistics is explained by the independent variables.

All of the coefficient estimates are statistically significant except for the coefficient on the percentage of pooled cost amounts and slope in the Turnbull WTP t-statistic model and the height of the tail in the $WTP1$ model. The lack of statistical significance in the Turnbull WTP t-statistic model is expected since pooling smooths the dependent variable and fat tails are not part of the Turnbull WTP calculation. The height of the tail does not matter in the $WTP1$ model because the WTP estimate is the cost amount where the probability of a yes response is 50% which is not sensitive to the tail. Note also that we do not include the fat tail variable (pctyes2) in the Turnbull model. In a model that includes the fat tail, as the height of the tail increases by 0.10 units the Turnbull t-statistic increases by 2.4. A flat Turnbull function would have a t-statistic above 24. This also perversely causes the percentage of pooled bid amounts to have a negative and statistically significant ($p < 0.10$) effect on the Turnbull t-statistic.

As the percentage of pooled bids in each of the other models increases the t-statistics

[4] We have 53 clusters instead of 51 since Alberini et al. (1997) uses data from 3 different studies.

decrease. If pooling doubles from its mean of 21.6%, then the mean $WTP1$, mean $WTP2$, median $WTP3$ and linear $WTP4$ t-statistics will fall by 1.20, 1.52, 1.27 and 1.75, respectively. As the height of the tail doubles from its mean of 23.3% then the mean $WTP2$, median $WTP3$ and linear $WTP4$ t-statistics will fall by 1.56, 1.03 and 2.00, respectively.

As the absolute value of the slope of the tail increases (i.e., gets steeper) the t-statistic increases. If the slope doubles from its mean of 0.48, then the mean $WTP1$, mean $WTP2$, median $WTP3$ and linear $WTP4$ t-statistics will increase by 0.86, 0.81, 0.78 and 0.94, respectively. In each of the models an increase in the sample size increases the t-statistic. If the sample size doubles from its mean of 433 the t-statistics will increase by 2.60, 3.48, 2.49, 1.23 and 3.00 for the Turnbull $WTP0$, mean $WTP1$, mean $WTP2$, median $WTP3$ and linear $WTP4$ t-statistics, respectively.

We next consider the effects of non-monotonicities, fat/flat tails and sample size on the ratio of the width of the Krinsky-Robb confidence interval to the width of the Delta Method confidence interval (Table 9). We estimate models for mean $WTP1$ and mean $WTP2$. In the $WTP1$ model, the ratio increases with the height of the tail of the distribution and decreases with sample size. The ratio increases by 55% if the height of the tail doubles from the average and decreases by 21% if the sample size doubles from the average. The sample size that equates the width of the $WTP1$ confidence intervals is $n = 1100$. In the $WTP2$ model, the ratio increases with the height of the tail of the distribution and decreases with the slope of the tail and sample size. The ratio increases by 37% if the height of the tail doubles from the average, decreases by 12% if the slope steepens by twice the mean and decreases by 19% if the sample size doubles from the average. The sample size that equates the width of the $WTP2$ confidence intervals is

$n = 1900$. In summary, fat and flat tails cause the Krinsky-Robb confidence interval to widen relative to the Delta Method confidence interval and increases in the sample size cause them to converge.

Finally, we estimate the effects of non-monotonicities, fat/flat tails and sample size on the asymmetry of the Krinsky-Robb confidence interval confidence interval (Table 10). In the $WTP1$ model, the ratio of the upper tail to the lower tail increases with the height of the tail of the distribution. The ratio increases by 129% if the height of the tail doubles from the average. In the $WTP2$ model, the ratio increases with the number of non-monotonicities and the height of the tail of the distribution and decreases with the slope of the tail and the sample size. The ratio increases by 48% if the percentage of non-monotonicities doubles from the average, increases by 100% if the height of the tail doubles from the average, and decreases by 32% if the sample size doubles from the average. In summary, fat and flat tails cause the Krinsky-Robb confidence interval to widen relative to the Delta Method confidence interval and increases in the sample size causes them to converge.

Replication of Split-Sample Hypothesis Tests

Twenty-five of the 51 studies contain data sets that support split-sample WTP comparison tests and 16 of these studies allow for directional hypotheses tests. Six of the 16 studies allow for 1 test each, 2 studies support 2 tests each, 4 studies support 3 tests, 3 studies support 6 tests and 1 study supports 12 tests. In total there are 52 possible directional hypothesis tests. Seventeen of these tests, including 12 from a single study, are for differences in individual health risk, 9 are for the scope of the policy, 15 are for hypothetical bias, and 9 are for payment schedules.

The test for differences in individual health risk is $\partial WTP/\partial r > 0$, where r is the risk that would be avoided by purchase of a treatment or payment for a policy. A scope test is similar with $\partial WTP/\partial q > 0$, where q is an environmental good. A test for hypothetical bias concerns comparing actual, A, and hypothetical, H, payments for a good or service, with an expectation of $WTP^H > WTP^A$. A test for payment schedules involves differences in the amount of time, t, a fixed payment would be made, $\partial WTP/\partial t < 0$. Each of these tests is directional and one-sided t-tests for differences in means are appropriate (Cho et al. 2013). We conduct t-tests for differences in WTP estimates across treatments with the Turnbull $WTP0$, mean $WTP1$, mean $WTP2$, median $WTP3$ and linear mean $WTP4$ estimates: $t-statistic = \frac{WTP_X - WTP_Y}{\sqrt{se_X^2 + se_Y^2}}$, where X and Y are different treatments. Our focus here is on statistically significant differences in the WTP estimates and not the significance of the economic differences.

The hypotheses tests are presented in Table 11. Considering first the Turnbull and mean $WTP1$ tests with the samples that do not suffer from negative WTP, the mean p-value on the t-statistic is 44% larger for the parametric mean $WTP1$ estimates relative to the Turnbull $WTP0$ estimates. With these 34 tests, 38% of the differences in Turnbull $WTP0$ estimates are statistically different at the 99% confidence level, 6% at the 95% level, and 6% at the 90% level. In contrast, only 12% of the differences in mean $WTP1$ estimates are statistically different at the 99% confidence level, 9% at the 95% level, and 6% at the 90% level.

With the full sample of 52 tests, the mean p-value on the t-statistic is 63%, 83%, and 60% larger for the parametric mean $WTP2$, median $WTP3$, and linear $WTP4$ estimates relative to the Turnbull $WTP0$ estimates. Forty-eight percent of the differences in Turnbull $WTP0$ estimates are statistically different at the 99% confidence level, 4% at the 95% level, and 8% at the 90%

level. Twenty-seven percent of the differences in mean $WTP2$ estimates are statistically different at the 99% confidence level, 2% at the 95% level, and 8% at the 90% level. Twenty-one percent of the differences in median $WTP3$ estimates are statistically different at the 99% confidence level, 2% at the 95% level, and 12% at the 90% level. Twenty-seven percent of the differences in linear mean $WTP4$ estimates are statistically different at the 99% confidence level, 13% at the 95% level, and 17% at the 90% level. The Turnbull and linear probability model estimates find some level of statistical significance most often, 60% and 58% respectively, relative to the tests for differences in the mean $WTP2$ (37%) and median $WTP3$ (35%) estimates.[5]

The statistical significance of the differences in WTP estimates can be increased by pooling the data and constraining the marginal utility of income to be equal across treatments. Such a constraint is economically reasonable as there is no theoretical reason why the marginal utility of income should differ across treatments. But the constraint may not be statistically appropriate, especially at smaller sample sizes. Conduct of the hypothesis tests with pooled samples should proceed only after the constraint is not rejected statistically (for an example, see the Appendix).

There are several ways in which a parametric split-sample hypothesis test can be conducted. The first is to pool the samples and include a treatment dummy variable for the differences in treatments: $\Pr(yes) = 1/(1 + \exp(-(\alpha + \beta A + \gamma D))$, where D is a dummy variable equal to 0 for a base case scenario and 1 for a treatment. One test is for differences in the probability of a yes response to the single-bound question, $\gamma \gtrless 0$. This test may produce

[5] We have estimated models similar to Table 6 with the t-statistics from the hypotheses tests and do not find any statistically significant determinants.

higher t-statistics since the coefficient on the treatment dummy variable is not divided by the coefficient on the cost amount. Another test is for whether the willingness to pay estimates from a pooled logit model are statistically different. For the Hanemann mean WTP2 estimate this test is for differences in $WTP(D = 0) = (-1/\beta) \times ln\big(1 + exp(\alpha)\big)$ and $WTP(D = 1) = (-1/\beta) \times ln\big(1 + exp(\alpha + \gamma)\big)$. These tests may produce higher t-statistics on the difference in willingness to pay because the marginal utility of income is constrained across samples. This constraint decreases the standard error of willingness to pay and, in some cases, increases the difference in willingness to pay.

Thirty-three of 52 tests for mean $WTP2$ have $p < 0.10$ and are candidates for less onerous tests with the Delta Method. These tests are from 11 articles. Fifteen of the tests are for differences in individual health risk, 7 are for hypothetical bias, 6 are for different payment schedules, and 5 are scope tests. We find a statistically significant treatment dummy coefficient estimate in 17 of the 33 tests. We find statistically significant differences in mean $WTP2$ in 15 of the 33 tests. Five of the tests are statistically significant at the $p < 0.10$ level, 2 are statistically significant at the $p < 0.05$ level, and 8 are statistically significant at the $p < 0.01$ level. For 9 of the 15 tests, the constraint that the base and treatment slope coefficients are statistically equal is rejected. There is no theoretical reason for different slope coefficients since the marginal utility of income should be constant. But, behaviorally, it may be logical for survey respondents to be less responsive to the cost amount for larger health risks, larger scope levels, longer payment schedules and hypothetical, relative to real, scenarios.[6]

[6] We are currently conducting the convolutions test with the Krinsky-Robb simulations (Poe, Severance-Lossin, and Welsh 1994, Poe, Giraud, and Loomis 2005).

Conclusions

In this paper we have replicated nonparametric and parametric willingness to pay estimates from 120 single-bound data sets in 51 CVM studies. We find that willingness to pay estimates can be unreliable; i.e., in many cases, willingness to pay estimates vary significantly depending on the estimation approach. This variation is by design in the case of the Turnbull, which is a lower bound estimate most appropriate for applications such as natural resource damage assessment (Carson et al. 2003) and sensitivity analysis in benefit-cost analysis. Considering parametric willingness to pay estimates, we focus our attention on three often-used measures from Hanemann (1984, 1989). A significant portion of the mean WTP estimates that allow for negative willingness to pay in the logistic function are negative overall and many others are not statistically different from zero. The WTP estimates from the approach that truncates the logistic distribution at zero are four times larger than the more conservative mean WTP estimates. This difference makes it unclear which willingness to pay measure should be used in benefit-cost analysis.

We estimate standard errors and t-statistics for these WTP estimates and find that the Turnbull WTP estimates are measured much more precisely then the parametric WTP estimates. The Turnbull WTP average t-statistic is 56% higher than the zero truncated mean WTP t-statistic. We find that the number of non-monotonicities in the cost amounts and fat tails contribute to lowering t-statistics in the parametric WTP estimates. Small sample size also contributes to low t-statistics.

We identify and conduct 52 split-sample tests of directional hypotheses (e.g., scope, hypothetical bias) in the 120 data sets. With relatively small standard errors, the Turnbull WTP estimates are more likely to lead to a researcher failing to reject the null hypothesis relative to tests conducted with the parametric WTP estimates estimated from the logit. Sixty-percent of tests conducted with the Turnbull WTP estimates find statistically different WTP estimates compared to 37% for the truncated mean and 35% for the median WTP estimates. Fifty-eight percent of difference in means tests conducted with the linear probability model are statistically significant. Considering only those tests with positive WTP and the more conservative mean WTP estimate, statistically significant differences in Turnbull WTP estimates are almost twice as common as parametric WTP differences.

The results of these tests should not be taken as a meta-analysis on the validity of the contingent valuation method (Boyle and Bishop 2019). Lower p-values may be achieved with each of these data sets with appropriate statistical models or by inclusion of covariates (see Appendix). Our only goal is to determine if there are any differences in the directional hypothesis tests across estimation approaches. We find that there are and caution researchers who may be tempted to rely on a single WTP estimation approach. In particular, statistical tests based on the Turnbull WTP estimate may be misleading relative to tests based on parametric methods.

These results lead to two conclusions. The first involves the Turnbull WTP estimate. As stated before, the Turnbull WTP estimate should be considered appropriate only for limited uses, for example, natural resource damage assessment or as a lower bound in sensitivity analyses of WTP in benefit-cost analysis. Sole reliance on the Turnbull WTP estimate is less appropriate for conducting directional hypothesis tests when assessing the validity of the contingent valuation method.

The second conclusion is that efforts should be made to better estimate WTP and its standard errors in parametric models with single-bound question. Our meta-analysis finds that these problems are lessened and may disappear with larger sample sizes. While there is an already literature on bid design and empirical approaches to modelling the preponderance of zero WTP (Kristrom 1997), additional research could focus on methods to avoid negative WTP and reduce the fat tails and flat tails problems.

Two approaches have emerged in the literature to collect additional information from survey respondents and improve the estimation of willingness to pay. In the first approach, follow-up dichotomous choice questions have been used to increase statistical efficiency (Hanemann, Loomis and Kanninen 1991). Doubled-bounded referendum questions present a follow-up question where respondents who vote for a policy at a tax amount are asked the same question at a higher tax amount. Respondents who vote against the policy are asked the same question with a lower tax amount. The amount of willingness to pay information provided by the respondent is increased. For respondents who change their vote (e.g., for-against and against-for) willingness to pay is bounded between the two cost amounts. For respondents who vote against the policy in the first and follow-up question, the range of willingness to pay above zero is narrower. For respondents who vote yes to the first and follow-up questions, the lower bound of willingness to pay is higher and the lower bound and income/infinity bound narrows. While a number of studies continue to use the double-bounded approach, this approach has been found to be prone to starting point bias and incentive incompatibility (Whitehead 2002, 2004). Use of double-bounded questions must be conducted with the knowledge that increased efficiency is obtained at the risk of bias.

In the second, more recent, approach, follow-up dichotomous choice questions have been used to increase statistical efficiency but the cost amounts that follow the first question are not anchored to the first question and other attributes vary as in discrete choice experiments. Vossler, Doyon, and Rondeau (2012) develop theory to show that a sequence of binary choice questions format is incentive-compatible if respondents treat each scenario as independent. Giguere, Moore and Whitehead (2020) find that while single binary choice questions produce WTP estimates that do not pass scope tests, the efficiency of the WTP estimates in a sequence of binary choice questions leads to WTP estimates that do exhibit sensitivity to scope. Thus, this type of study design, which blurs the distinction between contingent valuation and discrete choice experiments (Haab, Lewis and Whitehead, 2022) can be used as a reliable and useful alternative to contingent valuation surveys that employ a single dichotomous choice question.

Table 1. Data Summary by Study (n=52)					
Data Summary by Study					
Variable	Label	Mean	Std Dev	Minimum	Maximum
Year	publication year	2005.12	7.65	1990	2022
US	1 if USA data	0.61	0.49	0	1
Donation	1 if donation payment vehicle	0.22	0.42	0	1
Mail	1 if mail/mailback survey	0.47	0.5	0	1
Inperson	1 if in-person contact survey	0.25	0.44	0	1
Lab	1 if lab survey	0.14	0.35	0	1
Phone	1 if phone contact/survey	0.14	0.35	0	1
Online	1 if online contact/survey	0.06	0.24	0	1
Students	1 if student sample	0.14	0.35	0	1
Public	1 if public good	0.71	0.46	0	1
Costs	number of cost amounts	7.51	3.65	3	21
Onetime	one-time payment	0.53	0.5	0	1
Years	payment years	7.94	11.47	1	30
MinCost	minimum cost	23.14	42.35	0.5	200
MaxCost	maximum cost	1032.65	3526.01	2.5	24000

Table 2. Data summary by data set (n=120)					
Variable	Label	Mean	Std Dev	Minimum	Maximum
Sample	sample size (n)	433.31	529.58	47	4361
Costs	number of cost amounts	6.72	3.4	3	21
n/costs	sample size per cost amount	70.53	74.4	7	396
Pctpool	percent non-monotonicities	0.22	0.20	0	0.67
Sminbid	standardized minimum bid	0.10	0.12	0.00	0.67
Sbid1	standardized bid1	0.56	0.18	0.06	0.88
Sbid2	standardized bid2	0.88	0.20	0.25	1
Pctyes1	percent yes at Sbid1	0.35	0.16	0.03	0.82
Pctyes2	percent yes at Sbid2 (Fat tail)	0.23	0.16	0	0.74
Flat tail	standardized Kriström |slope|	0.48	0.52	0.01	4.02

Table 3. Willingness to Pay Estimates						
	Full Sample			Negative Mean WTP1 deleted		
Variable	Mean	SD	Cases	Mean	SD	Cases
Turnbull WTP0	331.50	764.30	120	331.66	730.32	99
Mean WTP1	233.70	1362.39	120	409.36	935.13	99
Mean WTP2	1099.73	6137.59	120	617.21	1438.25	99
Median WTP3	464.85	2473.38	119	547.49	2718.09	98
Linear WTP4	1087.76	6293.45	120	585.81	1404.57	99

Table 4. WTP ratios

	Ratio	SD	Sample
Mean WTP2 / Turnbull WTP0	1.73	1.16	99
Mean WTP2 / Mean WTP1	4.06	21.00	99
Mean WTP2 / Median WTP3	2.51	2.07	98
Mean WTP2/ Linear WTP4	1.09	0.11	99

Table 5. WTP t-statistics (Delta Method)

	t-statistic	SD	Sample
Turnbull WTP0	11.94	6.55	120
Mean WTP1	6.14	4.89	99
Mean WTP2	7.68	4.96	120
Median WTP3	5.29	3.81	119
Linear WTP4	8.72	5.84	120

Table 6. Krinsky-Robb to Delta Method Ratios of Confidence Intervals

	Mean	Min	Max	Sample
Mean WTP1	1.41	1.01	5.51	97
Mean WTP2	1.65	1.00	6.32	118

Table 7. Krinsky-Robb Confidence Interval Asymmetries

	Mean	Min	Max	Sample
Mean WTP1	1.50	0.49	6.94	67
Mean WTP2	2.66	1.19	8.41	108

Table 8. Determinants of t-statistics for WTP estimates

	t-statistic									
	WTP0		WTP1		WTP2		WTP3		WTP4	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Constant	9.65	6.44	3.56	3.32	7.45	7.21	5.56	6.65	8.52	7.57
Non-monotonicities (Pctpool)	0.32	0.09	-5.56	-2.99	-7.05	-4.48	-5.91	-3.51	-8.13	-4.70
Fat tail (Pctyes2)			-0.94	-0.51	-6.67	-3.56	-4.43	-3.11	-8.60	-4.18
Flat tail	-0.79	-1.02	1.81	3.49	1.69	3.48	1.62	2.96	1.97	3.65
Study sample size	0.006	3.32	0.008	8.01	0.006	4.36	0.003	3.15	0.007	4.09
Sample size	120		99		120		119		120	
R^2	0.24		0.47		0.58		0.34		0.61	
F-statistic (df)	12.52 (3)		20.84 (4)		38.91 (4)		14.61 (4)		44.24 (4)	

Table 9. Determinants of Krinsky-Robb to Delta Method Ratios

	Ratio			
	WTP1		WTP2	
	Coeff.	t-stat	Coeff.	t-stat
Constant	0.99	5.39	1.40	5.65
Non-monotonicities (Pctpool)	0.05	0.13	0.98	1.61
Fat tail (Pctyes2)	2.37	2.91	1.59	1.78
Flat tail	0.00002	0.00	-0.26	-1.79
Study sample size	-0.00049	-2.74	-0.00045	-2.34
Sample size	97		118	
R^2	0.27		0.47	
F-statistic (df)	8.64 (4)		20.84 (4)	

Table 10. Determinants of Krinsky-Robb Asymmetries

	Asymmetry			
	WTP1		WTP2	
	Coeff.	t-stat	Coeff.	t-stat
Constant	-0.90	-0.25	1.87	6.36
Non-monotonicities (Pctpool)	0.12	0.17	2.21	2.56
Fat tail (Pctyes2)	5.54	4.78	4.29	4.00
Flat tail	0.34	0.00	-0.41	-1.53
Study sample size	-0.00022	-1.14	-0.00074	-2.66
Sample size	67		118	
R^2	0.55		0.34	
F-statistic (df)	18.58 (4)		13.12 (4)	

Table 11. Average p-values and proportion of tests with t-statistics above the critical t-value in a one-tailed test

			Significance Level			
	Number of tests	Mean p-values	99%	95%	90%	90%+
Turnbull WTP0 (WTP1 > 0)	34	0.146	38%	6%	6%	50%
Logit Mean WTP1 (WTP1 > 0)	34	0.211	12%	9%	6%	26%
Turnbull WTP0	52	0.120	48%	4%	8%	60%
Logit Mean WTP2	52	0.195	27%	2%	8%	37%
Logit Median WTP3	51	0.226	21%	2%	12%	35%
Linear Mean WTP4	52	0.192	27%	13%	17%	58%

Appendix

Pooled Data Model Hypotheses Tests

There are several ways in which a parametric split-sample hypothesis test can be conducted. The first is to pool the samples and include a dummy variable for the differences in treatments: $\Pr(yes) = 1/(1 + \exp(-(\alpha + \beta A + \gamma D))$, where D is a dummy variable equal to 0 for a base case scenario and 1 for a treatment. One test is for differences in the probability of a yes response to the single-bound question. Another test is for whether the willingness to pay estimates from the logit model are statistically different. For the Hanemann mean WTP1 estimate this test is for differences in $WTP(D = 0) = -\alpha/\beta$ and $WTP(D = 1) = -(\alpha + \gamma)/\beta$. For the Hanemann mean WTP2 estimate this test is for differences in $WTP(D = 0) = (-1/\beta) \times (ln(\alpha))$ and $WTP(D = 1) = (-1/\beta) \times (ln(\alpha + \gamma))$. Each of these tests will produce higher t-statistics as the marginal utility of income is constrained across samples.

Berrens et al. (1996) conduct a scope test with samples of 162 and 167. With the split sample data models we find insensitivity to scope in the mean WTP1 estimates (t = 1.25) and the WTP2 estimates (t = 1.22) with the Delta Method standard errors. In pooled data models the restriction on equality between the slope coefficients (the marginal utility of income) cannot be rejected ($\chi^2 = 1.58, df = 1$). With this constraint imposed, the t-statistics on the scope dummy variable in the logistic regression models are statistically significant at the p=0.05 level in a one-tailed test: linear bid (t=1.65) and logged bid (t=1.83). The t-statistic on the bid amount in the linear probability model is t=1.64 (p=0.10). We find that the data are sensitive to scope at the p=0.10 level with the mean WTP1 estimate (t=1.56), the mean WTP2 estimate (t=1.60), the median WTP3 estimate (t=1.61) and the linear WTP4 estimate (t=1.48).

BP/Deepwater Horizon CVM Study

Much of the data analyzed in this paper suffers from small sample sizes and, perhaps, research budgets that preclude extensive use of focus groups and pretesting. However, even the best contingent valuation studies suffer from these problems. A team of over twenty prominent economists and social scientists conducted a study for the National Oceanic and Atmospheric Administration to estimate the lost total value due to the 2010 Gulf of Mexico oil spill (Bishop et al. 2017). The study is state-of-the-art and has a large sample ($n = 3656$). There were five cost amounts ranging from \$15 to \$435. The percentage of votes for the policy fell monotonically from a high of 52% to a low of 24% in the "small injury" treatment and 58% to 28% in the "large injury" treatment. Analysis of these data suggests that it suffers from all of the problems described above except non-monotonicity.

Willingness to pay for the base and large scenario with the Turnbull WTP estimates are \$132 (5.38) and \$152 (5.65), respectively. The t-statistic for the difference in means test is 2.55 ($p < 0.01$). The logit models for the split sample small and large scenarios and a pooled model with a dummy variable for the large scenario are presented below:

	Base		Large		Pooled	
	Estimate	t ratio	Estimate	t ratio	Estimate	t ratio
Constant	-0.043	-0.58	0.144	1.94	-0.052	-0.83
Cost	-0.0029	-8.28	-0.0028	-8.29	-0.0028	-11.72
Large					0.205	2.96
Sample	1833		1823		3656	
χ^2	73.54		72.87		154.85	
Pseudo-R^2	0.030		0.029		0.032	

Each of the cost coefficients is statistically different from zero and the cost coefficient is not statistically different across treatments. To conduct this test we estimate two additional pooled models with base and scope dummy variables and no constant. In the first model we estimate two cost amount coefficients and in the second we constrain the cost coefficient to be equal across treatments. The likelihood ratio test indicates the constraint is appropriate (χ^2=0.04 [1 df]).

Since the constant in the base model is negative the $WTP1$ estimate is negative so we proceed with estimating mean $WTP2$ values. Given the large sample sizes the WTP estimates are very efficient but the differences across scenarios is small.

	Turnbull		Split Sample Models		Pooled Model	
	WTP0	SE	WTP2	SE	WTP2	SE
Small	132.36	5.38	232.44	20.33	235.44	15.83
Large	152.25	5.65	275.82	24.32	272.45	18.10

The difference in WTP is statistically significant with the Turnbull estimates at the p = 0.0054 level (t=2.55). The differences in WTP are statistically significant with the split sample and pooled model estimates at the p = 0.086 level (t=1.37) and p = 0.062 (t=1.54).[7]

However, these mostly happy results may be a situation where estimates are statistically different rather than economically different due to large sample sizes. Randomly selecting observations so that the sample size is a more typical $n = 1096$ increases the standard error on the scope effect coefficient in the pooled model so that the coefficient is no longer statistically different from zero in a one-tailed test at the $p = 0.10$ confidence level ($t = 1.15$). The lesson is that large sample sizes are needed with single bound valuation questions.

[7] The t-statistics from the Wald command and that from the formula are identical in the split-sample models. But, the Wald command produces a t-statistic of t=2.91 (p=0.0018) with the pooled model. We present the t-statistic from the formula above but have not determined the source of the difference.

References

Arrow, Kenneth, Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, and Howard Schuman. "Report of the NOAA panel on contingent valuation." Federal register 58, no. 10 (1993): 4601-4614.

Bengochea-Morancho, Aurelia, Ana Ma Fuertes-Eugenio, and Salvador del Saz-Salazar. "A comparison of empirical models used to infer the willingness to pay in contingent valuation." Empirical Economics 30 (2005): 235-244.

Bishop, Richard C., and Thomas A. Heberlein. "Measuring values of extramarket goods: Are indirect measures biased?." American Journal of Agricultural Economics 61, no. 5 (1979): 926-930.

Bishop, Richard C., Kevin J. Boyle, Richard T. Carson, David Chapman, W. Michael Hanemann, Barbara Kanninen, Raymond J. Kopp et al. "Putting a value on injuries to natural assets: The BP oil spill." Science 356, no. 6335 (2017): 253-254.

Bishop, Richard C., and Kevin J. Boyle. "Reliability and validity in nonmarket valuation." Environmental and Resource Economics 72 (2019): 559-582.

Carson, Richard. Contingent Valuation: A Comprehensive Bibliography and History. Edward Elgar Publishing, 2012.

Carson, Richard T., Nicholas E. Flores, Kerry M. Martin, and Jennifer L. Wright. "Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods." Land Economics (1996): 80-99.

Carson, Richard T., Robert C. Mitchell, Michael Hanemann, Raymond J. Kopp, Stanley Presser, and Paul A. Ruud. "Contingent valuation and lost passive use: damages from the Exxon Valdez oil spill." Environmental and resource economics 25 (2003): 257-286.

Carson, Richard T., and Theodore Groves. "Incentive and informational properties of preference questions." Environmental and Resource Economics 37, no. 1 (2007): 181-210.

Carson, Richard T., Theodore Groves, and John A. List. "Consequentiality: A theoretical and experimental exploration of a single binary choice." Journal of the Association of Environmental and Resource Economists 1, no. 1/2 (2014): 171-207.

Cho, Hyun-Chul, and Shuzo Abe. "Is two-tailed testing for directional research hypotheses tests legitimate?" Journal of Business Research 66, no. 9 (2013): 1261-1266.

Giguere, Christopher, Chris Moore, and John C. Whitehead. "Valuing hemlock woolly adelgid control in public forests: Scope effects with attribute nonattendance." Land Economics 96, no. 1 (2020): 25-42.

Haab, Tim, Lynne Y. Lewis, and John Whitehead. "State of the art of contingent valuation." Oxford Research Encyclopedia of Environmental Economics, James R. Kahn, editor, Oxford University Press (2022).

Haab, Timothy C., and Kenneth E. McConnell. "Referendum models and negative willingness to pay: alternative solutions." Journal of Environmental Economics and Management 32, no. 2 (1997): 251-270.

Haab, Timothy C., and Kenneth E. McConnell. Valuing environmental and natural resources: the econometrics of non-market valuation. Edward Elgar Publishing, 2002.

Hanemann, W. Michael. "Welfare evaluations in contingent valuation experiments with discrete responses." American Journal of Agricultural Economics 66, no. 3 (1984): 332-341.

Hanemann, W. Michael. "Welfare evaluations in contingent valuation experiments with discrete response data: reply." American Journal of Agricultural Economics 71, no. 4 (1989): 1057-1061.

Hanemann, Michael, John Loomis, and Barbara Kanninen. "Statistical efficiency of double-bounded dichotomous choice contingent valuation." American Journal of Agricultural Economics 73, no. 4 (1991): 1255-1263.

Hanemann, Michael, and Barbara Kanninen. "The Statistical Analysis of Discrete-Response CV Data." Chapter 11 in Valuing environmental preferences: theory and practice of the contingent valuation method in the US, EU, and developing countries (2001): 302.

Hole, Arne Risa. "A comparison of approaches to estimating confidence intervals for willingness to pay measures." Health Economics 16, no. 8 (2007): 827-840.

Johnston, R.J., Boyle, K.J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T.A., Hanemann, W.M., Hanley, N., Ryan, M., Scarpa, R. and Tourangeau, R., 2017. Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, *4*(2), pp.319-405.

Kriström, Bengt. "A non-parametric approach to the estimation of welfare measures in discrete response valuation studies." Land Economics 66, no. 2 (1990): 135-139.

Kriström, Bengt. "Spike models in contingent valuation." American Journal of Agricultural Economics 79, no. 3 (1997): 1013-1023.

Loomis, John B. "Contingent valuation using dichotomous choice models." Journal of Leisure Research 20, no. 1 (1988): 46-56.

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

Park, Timothy, and John Loomis. "Comparing models for contingent valuation surveys: Statistical efficiency and the precision of benefit estimates." Northeastern Journal of Agricultural and Resource Economics 21, no. 2 (1992): 170-176.

Parsons, George R., and Kelley Myers. "Fat tails and truncated costs in contingent valuation: An application to an endangered shorebird species." Ecological Economics 129 (2016): 210-219.

Poe, Gregory L., Eric K. Severance-Lossin, and Michael P. Welsh. "Measuring the difference (X—Y) of simulated distributions: A convolutions approach." American Journal of Agricultural Economics 76, no. 4 (1994): 904-915.

Poe, Gregory L., Kelly L. Giraud, and John B. Loomis. "Computational methods for measuring the difference of empirical distributions." American Journal of Agricultural Economics 87, no. 2 (2005): 353-365.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. "Truth in consequentiality: theory and field evidence on discrete choice experiments." American Economic Journal: Microeconomics 4, no. 4 (2012): 145-171.

Whitehead, John C. "Incentive incompatibility and starting-point bias in iterative valuation questions." Land Economics 78, no. 2 (2002): 285-297.

Whitehead, John C. "Incentive incompatibility and starting-point bias in iterative valuation questions: reply." Land Economics 80, no. 2 (2004): 316-319.