Department of Economics Working Paper

Number 25-11 | November 2025

A Meta-Analysis of SBC Contingent Valuation: Willingness to Pay Estimates, Determinants of Reliability and Replication of 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

Department of Economics
Appalachian State University
Boone, NC 28608
Phone: (828) 262-2148
Fax: (828) 262-6105
www.business.appstate.edu/economics

A Meta-Analysis of SBC Contingent Valuation: Willingness to Pay Estimates, Determinants of Reliability and Replication of 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

November 14, 2025

Abstract. The single binary choice (SBC) question format, commonly used in contingent valuation studies and modeled as a hypothetical referendum, is considered incentive compatible when paired with a coercive payment vehicle and a consequential survey. Despite its dominance in the field, the SBC format yields limited information, which can result in imprecise and unreliable estimates of willingness to pay (WTP). This chapter explores the limitations of SBC using a meta-analysis dataset originally compiled by Lewis, Richardson, and Whitehead (2024) for nonparametric WTP estimation. We extend their work by analyzing parametric WTP estimates and comparing them with nonparametric Turnbull and adjusted Kriström estimates. Our results show that parametric WTP can differ significantly from the Turnbull nonparametric estimate, and that confidence intervals derived from parametric models are often wider than those from non-paramteric WTP estimates. In a meta-regression, we find that the inefficiency of SBC decreases with data quality. We illustrate the importance of these issues with a replication of directional split-sample tests from the meta-data. Compared to parametric WTP estimates, tests using Turnbull and adjusted Kriström estimates are more likely to detect statistically significant differences in WTP, underscoring the importance of robustness tests with alternative WTP estimates.

Acknowledgements: This paper has benefited from presentation at the University of Wyoming, the University of Georgie, an Association of Environmental and Resource Economists (AERE) session at the Southern Economic Association meetings in 2024 and the AERE Summer Conference in 2025. The authors thank Christian Vossler for a number of helpful comments.

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). Following the introduction of the dichotomous choice response format by Bishop and Heberlein (1979), SBC (SBC) contingent valuation questions became the preferred question format. SBC questions present a survey respondents with a single cost and yes/no answer categories 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 has evolved to a for/against vote in the context of a policy referendum.

A number of influential publications have led to dominance of the SBC question format in the CVM literature. Hanemann (1984) developed the indirect utility theory to support the use of SBC 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 referendum question with a coercive payment

[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 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."

vehicle is incentive compatible. Carson, Groves and List (2017) conduct an experimental test of the incentive compatibility of the SBC question with consequentiality to bolster those claims. Finally, in an article on best practice recommendations for stated preference studies to support decision making, Johnston et al. (2017) recommend the use of the SBC 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 SBC question is the "gold standard" for value elicitation. And yet, the data that results from surveys that employ SBC questions are often problematic.

Econometric approaches to estimation of willingness to pay (WTP) with SBC 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 SBC 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 SBC 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 (Hanemann 1984, Haab and McConnell 1997). Hanemann (1989) provides a formula for estimating WTP that arbitrarily—and incorrectly in a statistical sense--eliminates this negative

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

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 extremely sensitive to the assumed logged distribution, and can be undefined (Haab and McConnell 2002). The nonparametric Turnbull (Haab and McConnell 1997) assumes a lower bound on WTP at zero and masses the upper portion of the distribution at the highest cost amount. 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. The Kriström nonparametric estimator extends the Turnbull by estimating the slope between cost amounts with linear interpolation and avoiding truncation at the highest bid by estimating a choke price (Kriström 1990). 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 aggregate 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 slope over the entire range of cost amounts. Beyond the problem of a lack of theoretical validity in pair-wise comparisons of cost amounts that exhibit non-monotonicities, this empirical issue 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 shape of the upper tail all together, resulting in the Turnbull failing to provide a point estimate of expected WTP and instead only providing a lower bound on expected WTP that is sensitive to the set of bids offered in the survey. 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 (i.e., choke price). The linear probability model estimates the choke price 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 probability of a yes response is relatively flat at two or more of the highest cost amounts (Lewis, Richardson and Whitehead 2024). Flat tails lead to less precise WTP estimates.

In this chapter we investigate these problems with SBC data 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 SBC data is increasing in the percentage of non-monotonicities over

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

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 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 smoothed nonparametric 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.[4]

In the next section we first review the theory and estimation of SBC 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.

SBC Contingent Valuation

Suppose a consumer has a willingness to pay for a change in the quality or quantity, q, of a private, public or quasi-public good, $WTP(\Delta q)$. A binary choice 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 often posed as a referendum vote for public goods, 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 private or quasi-

[4] This is to be expected as the Turnbull is not providing a point estimate of expected WTP, but rather providing an estimate of the lower bound on expected WTP, a point that is often confused in the literature.

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 behave strategically when a government or private business firm 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.

A number of WTP estimates have been developed Czajkowski et al. (2024). In this chapter we compare two nonparametric measures and four parametric measures of WTP. A summary of these WTP estimators is presented in Table 1.

The consumer will answer yes/for to the 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 expected WTP by assuming non-negative WTP: $WTP0 = \sum_j A_j \times [\Pr(yes_j) - \Pr(yes_{j+1})]$, where $\Pr(yes_{j-1}) = 0$ at the lowest cost amount. Essentially, the Turnbull calculates the sample proportion of respondents falling between cost amounts and assigns every between bid proportion a WTP equal to the 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 yes response of 100% when $A = 0$: $WTP1 = \sum_j \frac{1}{2}(A_j + A_{j+1}) \times [\Pr(yes_j) - \Pr(yes_{j+1})]$. Lewis, Richardson and Whitehead

(2024) find that the Kriström WTP estimates are susceptible to the fat tails problem and assess a correction proposed by Richardson and Lewis (2022), termed the adjusted Kriström. The choke price, A_C where $\Pr(yes_C) = 0 \mid A_C$, is estimated by using the slope of the bid function from the linear probability model, $\varphi < 0$ below, and estimating the choke price from the highest cost amount, A_k, $A_C = A_k - \varphi \times \Pr(yes_k)$.

Hanemann (1984) provided theoretical justification for SBC contingent valuation. Beginning with a linear utility function, Hanemann shows that the logit (and probit) model relies on the notion that median 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 $WTP3 = -\alpha/\beta$. This is sometimes referred to as "mean" WTP.

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 $WTP3 < 0$. Hanemann (1989) proposed a widely used correction that truncates the negative portion of the WTP distribution: $WTP4 = (-1/\beta) \times \ln\left(1 + \exp(\alpha)\right)$. This is sometimes referred to as the "truncated mean" WTP estimate. However, as Haab and McConnell (2002) note, the Hanemann procedure does not produce a statistically valid WTP estimate as the arbitrary truncation results in a WTP distribution that does not integrate to one. A logged cost amount model, $\Pr(yes) = 1/(1 + \exp(-(\gamma + \delta lnA))$ has also been used to solve the "negative WTP" problem. The resulting WTP estimate is, $WTP5 = exp\left(-(\gamma/\delta)\right)$. This is also sometimes referred to as "median WTP."

Finally, willingness to pay can also be estimated from a linear probability model: $\Pr(yes) = \theta + \varphi A$, although, to our knowledge, this has not appeared in the peer-reviewed literature (Loomis 1988). Intuitively, WTP is the triangle under the regression line: $WTP6 = 0.5 \times \theta \times (-\theta/\varphi)$. The area of this WTP triangle is the parametric version of the nonparametric adjusted Kriström WTP estimate.

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 and adjusted Kriström WTP estimates are estimated in MS Excel. The logit and OLS regression models of cost on the yes/no responses are $\Pr(yes) = 1/(1 + \exp(-(2.80 - 0.14 \times A)))$ and $\Pr(yes) = 0.95 - 0.021 \times A$, respectively. The willingness to pay estimates (with standard errors in parentheses) are \$18 (0.57), 21 (0.40), \$20 (0.64), \$21 (0.59), \$18 (0.61) and \$21 (0.55) from the Turnbull ($WTP1$), adjusted Kristrom ($WTP2$), Hanemann logit ($WTP3$-$WTP5$) and linear probability ($WTP6$) models, respectively. The willingness to pay estimates are not statistically different across valuation method. We consider these data "reliable" since they produce the same WTP amount regardless of estimation method. Unfortunately, as we will see, real world SBC data sets are not so well-behaved as textbook data and do not produce reliable WTP estimates.

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 needed to estimate WTP (Lewis, Richardson and Whitehead 2024). In addition to these studies, Lewis, Richardson and Whitehead (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 2. 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 3). 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 when estimating WTP. In order to make the cost amounts comparable across studies as independent variables, for each individual study we divide each cost amount by the maximum cost amount so that the standardized cost amounts can range from zero to one. The mean of the standardized minimum cost amount is 0.10 (i.e., 10% of the highest cost amount). 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 cost amounts for one of the cost amounts used to calculate this slope. The mean of the standardized low cost amount in the slope (Sbid1) is 0.56 and the mean of the standardized high cost amount 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 costs is 0.48 with a range of 0.01 to 4.02.

WTP Estimates

We construct the Turnbull $WTP1$ and adjusted Kristrom $WTP2$ estimates in MS Excel and the Hanemann mean $WTP3$, truncated mean $WTP4$, median $WTP5$ and linear probability $WTP5$ estimates from logit and linear probability models for each of the 120 data sets (Table 4). One of the median $WTP5$ estimates approached infinity so it is dropped from the data summary. As expected, the Turnbull lower bound $WTP0$ is 332, lower than all of the other WTP estimates except mean $WTP3$ for which 18% of the values are negative. The third lowest estimate is the Hamemann median ($WTP5 = 465$) followed by the adjusted Kristrom ($WTP2 = 533$). The means of the remaining WTP estimates are greater than 1000. The truncated mean $WTP4$ and linear $WTP6$ estimates are, not surprisingly, very similar since both estimates disregard the negative portion of the WTP distribution and are sensitive to the tail of the distribution. We next delete the WTP estimates for which the Mean $WTP3$ estimate is negative. Of the remaining 99 samples, the WTP estimates are much closer in magnitude with mean WTP estimates ranging from 332 ($WTP1$) to 617 ($WTP4$). Much of the variability in WTP estimates is due to data sets that produce negative $WTP3$ estimates. We conclude that the WTP estimates from many of these data sets suffer from unreliability.

Confidence Intervals

We estimate the standard errors of the Turnbull $WTP1$ estimates with the formula found in Haab and McConnell (p. 75, 2002) and the adjusted Kriström $WTP2$ standard errors are estimated following Boman, Bostedt and Kriström (1999). 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 $WTP1$ and adjusted Kriström 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 standard errors from the textbook data), as well as the facts that the nonparametric survival functions are smoothed when non-monotonicities are encountered and do not have a fat or flat tails. 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 $WTP1$ and Kriström $WTP2$ estimates are statistically significant. In contrast, 13% (n=13) of the non-negative Hanemann mean $WTP3$ 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 $WTP5$ estimates are not statistically different from zero and 2.5% of the truncated mean $WTP4$ and $WTP6$ 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 CVM 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.

To estimate the Krinsky-Robb 95% confidence intervals for the Hanemann mean $WTP3$ and $WTP4$ estimates, we simulate one million WTP estimates and trim the lowest and highest 2.5% values. Krinsky-Robb confidence intervals are significantly wider than Delta Method confidence intervals (Table 6). Of $n = 99$ estimates where mean *WTP3* 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 $WTP4$ 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 $WTP3$ Krinsky-Robb confidence intervals include zero. In contrast, only 18% of the Delta Method non-negative mean $WTP3$ confidence intervals include zero. Ten percent of the mean $WTP4$ Krinsky-Robb confidence intervals include zero. Only 3% of the Delta Method non-negative mean $WTP4$ 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

$Asymmetry = (U95 - WTP)/(WTP - L95)$, where $U95$ is the upper 95% Krinsky-Robb bound and $L95$ is the lower 95% Krinsky-Robb bound. The Krinsky-Robb asymmetry ratio for mean $WTP3$ is 1.50 with a range of 0.49 to 6.94 (Table 7). The Krinsky-Robb asymmetry ratio for mean $WTP4$ 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 linear regression models with the WTP t-statistics as the dependent variables (Table 8). The independent variables are the percentage of the number of pooled bids (non-monotonicities), the height (fat tail) and slope of the tail (flat tail), and study sample size. The standard errors are clustered at the study level.[5] 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 in the model with the Turnbull WTP t-statistic as the dependent variable, the fat tail slope in the Turnbull and adjusted Kristrom models, and the height of the tail in the Hanemann $WTP5$ model. The lack of statistical significance in the Turnbull WTP t-statistic model is expected since pooling smooths the dependent variable and flat tails do not inflate WTP standard errors. Note 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 each 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 cost

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

amounts to have a negative and statistically significant ($p < 0.10$) effect on the Turnbull WTP t-statistic. We find a similar fat tail result in the adjusted Kristrom model at the p=0.11 level. But, dropping this variable does not significantly affect the results. The height of the tail does not matter in the $WTP3$ 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.

As the percentage of pooled bids in each of the other models increases the t-statistics decrease. If pooling doubles from its mean of 21.6%, then the Kristrom $WTP2$, Hanemann mean $WTP3$, mean $WTP4$, median $WTP5$ and linear $WTP6$ t-statistics will fall by 1.61, 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 $WTP4$, median $WTP5$ and linear $WTP6$ 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-statistics increase. If the slope doubles from its mean of 0.48, then the adjusted Kriström $WTP2$, Hanemann mean $WTP3$, mean $WTP4$, median $WTP5$ and linear $WTP6$ t-statistics will increase by 1.25, 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, 4.32, 3.48, 2.49, 1.23 and 3.00 for the Turnbull $WTP0$, adjusted Kriström, Hanemann 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 for the $WTP3$ and $WTP4$ estimates (Table 9). In the $WTP3$ model, the ratio of the Krinsky-Robb to Delta Method interval increases with the height of the fat tail 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 $WTP3$ confidence intervals is $n = 1100$. In the $WTP4$ 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 for the Hanemann $WTP3$ and $WTP4$ estimates (Table 10). In the $WTP3$ model, the ratio of the upper tail to the lower tail increases with the height of the fat tail. The ratio increases by 129% if the height of the tail doubles from the average. In the $WTP4$ model, the ratio increases with the number of non-monotonicities and the height of the tail of the distribution and decreases with 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 tails cause asymmetries in the Krinsky-Robb 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 the 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 each of the WTP 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 economic significance.

The results of the hypotheses tests are presented in Table 11. The average p-value of all of the texts except for the adjusted Kriström $WTP2$ estimates indicates that, on the whole, the SBC data does not lead to statistically significant differences. Forty-two percent of the differences in Turnbull $WTP1$ estimates are statistically different at the 99% confidence level, 52% at the 95% level, and 60% at the 90% level. Fifty-seven percent of the differences in adjusted Kriström $WTP2$ estimates are statistically different at the 99% confidence level, 65% at the 95% level, and 73% at the 90% level. These passing rates are significantly higher than the passing rates for the parametric WTP hypothesis tests. Researchers should strongly reconsider each nonparametric estimators continued use in isolation.

Only three percent of the differences in the Hanemann mean $WTP3$ estimates are statistically different at the 99% confidence level, 21% at the 95% level, and 26% at the 90% level. This standard estimator already precludes hypothesis tests for 35% of the sample. Researchers should strongly reconsider its continued use in isolation

Fifteen percent of the differences in Hanemann truncated mean $WTP4$ estimates are statistically different at the 99% confidence level, 29% at the 95% level, and 37% at the 90% level.[6] Seventeen percent of the differences in median $WTP5$ estimates are statistically different at the 99% confidence level, 23% at the 95% level, and 35% at the 90% level. Seventeen percent

[6] We have also conducted the convolutions test with the Krinsky-Robb simulations (Poe, Severance-Lossin, and Welsh 1994, Poe, Giraud, and Loomis 2005). The results are similar to the results from the Delta Method. Seventeen percent of the differences in Hanemann truncated mean $WTP4$ estimates are statistically different at the 99% confidence level, 27% at the 95% level, and 38% at the 90% level.

of the differences in linear $WTP6$ estimates are statistically different at the 99% confidence level, 31% at the 95% level, and 38% at the 90% level.

There are several other 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: $\Pr(yes) = 1/(1 + \exp(-(\alpha + \beta A + \omega D))$, 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 SBC question, $\omega \gtrless 0$. This test may produce 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 $WTP4$ estimate this test is for differences in $WTP(D = 0) = (-1/\beta) \times ln\big(1 + exp(\alpha)\big)$ and $WTP(D = 1) = (-1/\beta) \times ln\big(1 + exp(\alpha + \omega)\big)$. 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 $WTP4$ 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 $WTP4$ in 15 of the 33 tests. Five of the tests are statistically significant at the 90% level, 2 are statistically significant at the 95% level, and 8 are statistically significant at the 99% 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.

Conclusions

In this chapter we have replicated nonparametric and parametric willingness to pay estimates from 120 SBC data sets in 51 CVM studies. We find that willingness to pay estimates can 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) and the linear probability model. 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 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 in the 120 data sets. With relatively small standard errors, the Turnbull and adjusted Kriström 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. Sixty-percent and 73% of tests conducted with the Turnbull and the adjusted Kristrom WTP estimates find statistically different WTP estimates compared to 37% for the truncated mean WTP, 35% for the median WTP and 38% for the linear WTP estimates.

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. Our primary goal is to determine if there are any differences in the directional hypothesis tests across WTP estimation approaches. We find that there are and caution researchers who may be tempted to rely on a single WTP estimation approach without robustness checks. In particular, statistical tests based solely on the nonparametric WTP estimates may be particularly misleading.

These results lead to the conclusion that efforts should be made to better estimate WTP and its standard errors in parametric models with the SBC question. Our meta-analysis finds that these problems are lessened and may disappear with larger sample sizes. While there is an already large literature on bid design and empirical approaches to modelling the preponderance of zero WTP values (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 discrete 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 SBC questions produce WTP estimates that do not pass scope tests, the increased 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 binary choice question.

More research is needed to determine if these types of CVM data improve the reliability of WTP across estimation methods, decrease WTP standard errors, equate Delta Method and Krinsky-Robb confidence intervals and lead to WTP estimates that are more likely to pass validity tests.

Table 1. Willingness to pay estimates	
Estimator	
Turnbull	$WTP1 = \sum_j A_j \times [\Pr(yes_j) - \Pr(yes_{j+1})]$
Adjusted Kriström	$WTP2 = \sum_j (1/2) \times (A_j + A_{j+1}) \times [\Pr(yes_j) - \Pr(yes_{j+1})]$, $\Pr(yes) = 1 \mid A = 0, A_C = A_k - \beta \times \Pr(yes_k)$
Hanemann (1984) "mean"	$WTP3 = -\alpha/\beta$
Hanemann (1989) "truncated mean"	$WTP4 = (-1/\beta) \times \ln(1 + \exp(\alpha))$
Hanemann (1984) "median"	$WTP5 = \exp(-\gamma/\delta)$
Linear probability model	$WTP6 = (1/2) \times \theta \times (-\theta/\varphi)$
Notes:	

Table 2. 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 3. 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 4. Willingness to Pay Estimates						
	Full Sample			Negative WTP3 deleted		
Variable	Mean	SD	Cases	Mean	SD	Cases
Turnbull (WTP1)	331.50	764.30	120	331.66	730.32	99
Adjusted Kriström (WTP2)	532.77	1207.03	120	539.93	1167.51	99
Hanemann (1984) "mean" (WTP3)	233.70	1362.39	120	409.36	935.13	99
Hanemann (1989) "truncated mean" (WTP4)	1099.73	6137.59	120	617.21	1438.25	99
Hanemann (1984) "median" (WTP5)	464.85	2473.38	119	547.49	2718.09	98
Linear probability model (WTP6)	1087.76	6293.45	120	585.81	1404.57	99

Table 5. WTP t-statistics (Delta Method)			
	t-statistic	SD	Sample
Turnbull (WTP1)	11.94	6.55	120
Adjusted Kriström (WTP2)	14.83	7.40	120
Hanemann (1984) “mean” (WTP3)	7.68	4.96	120
Hanemann (1989) “truncated mean” (WTP4)	5.29	3.81	119
Hanemann (1984) “median” (WTP5)	8.72	5.84	120
Linear probability model (WTP6)	6.14	4.89	99

Table 6. Krinsky-Robb to Delta Method Ratios of Confidence Intervals				
	Mean	Min	Max	Sample
Hanemann (1984) “mean” (WTP3)	1.41	1.01	5.51	97
Hanemann (1989) “truncated mean” (WTP4)	1.65	1.00	6.32	118

Table 7. Krinsky-Robb Confidence Interval Asymmetries				
	Mean	Min	Max	Sample
Hanemann (1984) “mean” (WTP3)	1.50	0.49	6.94	67
Hanemann (1989) “truncated mean” (WTP4)	2.66	1.19	8.41	108

Table 8. Determinants of t-statistics for WTP estimates

	t-statistic											
	WTP1		WTP2		WTP3		WTP4		WTP5		WTP6	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Constant	9.65	6.44	9.61	7.23	7.45	7.21	5.56	6.65	8.52	7.57	3.56	3.32
Non-monotonicities (Pctpool)	0.32	0.09	-7.50	-2.51	-7.05	-4.48	-5.91	-3.51	-8.13	-4.70	-5.56	-2.99
Fat tail (Pctyes2)			5.46	1.60	-6.67	-3.56	-4.43	-3.11	-8.60	-4.18	-0.94	-0.51
Flat tail	-0.79	-1.02	2.60	2.26	1.69	3.48	1.62	2.96	1.97	3.65	1.81	3.49
Study sample size	0.006	3.32	0.010	5.70	0.006	4.36	0.003	3.15	0.007	4.09	0.008	8.01
Sample size	120		120		120		119		120		99	
R^2	0.24		0.57		0.58		0.34		0.61		0.47	
F-statistic (df)	12.52 (3)		37.43 (4)		38.91 (4)		14.61 (4)		44.24 (4)		20.84 (4)	

Note: Turnbull (WTP1), Adjusted Kriström (WTP2), Hanemann (1984) "mean" (WTP3), Hanemann (1989) "truncated mean" (WTP4), Hanemann (1984) "median" (WTP5), Linear probability model (WTP6)

Table 9. Determinants of Krinsky-Robb to Delta Method Ratios				
	Ratio			
	WTP3		WTP4	
	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			
	WTP3		WTP4	
	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%
Turnbull (WTP1)	52	0.120	42%	52%	60%
Adjusted Kriström (WTP2)	34	0.211	3%	21%	26%
Hanemann (1984) "mean" (WTP3)	52	0.195	15%	29%	37%
Hanemann (1989) "truncated mean" (WTP4)	51	0.226	17%	23%	35%
Hanemann (1984) "median" (WTP5)	52	0.192	17%	31%	38%
Linear probability model (WTP6)	52	0.080	57%	65%	73%

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.

Boman, Mattias, Göran Bostedt, and Bengt Kriström. "Obtaining welfare bounds in discrete-response valuation studies: A non-parametric approach." Land Economics (1999): 284-294.

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 SBC." 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.

Czajkowski, Mikołaj, Ewa Zawojska, Norman Meade, Ronaldo Seroa da Motta, Mike Welsh, and Ramon Arigoni Ortiz. "On the inference about a willingness-to-pay distribution using contingent valuation data." Ecological Economics 222 (2024): 108207.

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.

Richardson, Leslie, and Lynne Lewis. "Getting to know you: individual animals, wildlife webcams, and willingness to pay for brown bear preservation." American Journal of Agricultural Economics 104, no. 2 (2022): 673-692.

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.