Department of Economics Working Paper

Number 24-04| February 2024

They doth protest too much, methinks: Reply to "Reply to Whitehead"

John Whitehead
Appalachian State University

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

They doth protest too much, methinks: Reply to "Reply to Whitehead"

John C. Whitehead
Department of Economics
Appalachian State University
Boone, NC 28608
whiteheadjc@appstate.edu

February 5, 2024

Abstract. Desvousges, Mathews and Train (2020) point out a mistake in my comment on their 2015 paper. When this mistake is corrected the conclusions drawn in my comment are unchanged. In addition, the authors claim that I make another 11 "mistakes". In this paper I argue that these "mistakes" are mostly fairly standard practice in the contingent valuation method. Desvousges, Mathews and Train misread and distort this literature. In addition, I place the comments and reply in the context of a larger debate over using the Contingent Valuation Method for Natural Resource Damage Assessment.

Introduction

Desvousges, Mathews and Train (2015) use the contingent valuation method (CVM) to conduct an adding-up test for willingness to pay estimated with the contingent valuation method (i.e., does $WTP_A + WTP_B = WTP_{A+B}$?). They use the nonparametric Turnbull estimator and find that the data do not pass the adding-up test. This suggests that the contingent valuation method lacks internal validity. In September 2016 I began writing a comment on this paper questioning the validity of the underlying data and implementation of the survey. The comment went through several rounds of review, was submitted, reviewed, revised and rejected at Land Econ[1], submitted, reviewed, revised and then withdrawn from *Economics E-Journal*, and submitted, reviewed and accepted for publication at *Ecological Economics* (Whitehead 2020).

Desvousges, Mathews and Train (Ecological Economics, 2020) replied to my comment by describing 12 mistakes. I agree that I made one of the mistakes on their list. I conducted an adding-up test by examining whether the confidence intervals for two willingness to pay estimates (the whole vs the sum of the parts) overlap. It is well-known that confidence intervals can overlap and yet the t-statistic for the test will indicate that the difference in means is statistically different. The mistake that I made was not checking the t-statistic. This is an embarrassing mistake. The worst part is that I taught this for many years to undergraduates in the business statistics course. I would ask the students not to make this mistake and I've made it in a

[1] The *Land Economics* referee said that there were serious problems with Desvousges, Mathews and Train's reply. The editor explained to me that the journal's policy was to publish only comment and reply pairs. Since the reply was rejected, the comment could not be published. This journal policy might create an incentive for a type of moral hazard. Authors who have comments written about their paper should write an unpublishable reply to kill the comment.

published journal article. I'm very embarrassed. Nevertheless, I bring the risk of further ridicule in an attempt to correct the record.

In the correction below, I point out another mistake that I made that concerns me almost as much as the t-statistic: *I used the wrong confidence intervals*. In Whitehead (2020) I used the confidence intervals from the Delta Method (a first-order Taylor Series expansion from the variance-covariance matrix) which are symmetric. It is well-known that 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 as in Desvousges, Mathews and Train (2015). Another approach that is common in the CVM literature is the Krinsky-Robb (KR) confidence intervals. These are based on a simulation from the variance-covariance matrix of the estimated parameters. Below I show that the KR confidence intervals using the Desvousges, Mathews and Train (2015) data are asymmetric and wide. The confidence intervals are so wide that the WTP for the sum of the parts lies within the confidence interval for the WTP for the whole, supporting the conclusions of Whitehead (2020).[2]

However, my biggest concern with Desvousges, Mathews and Train's "Reply to Whitehead" is that they do not take the problems with their own research very seriously. Desvousges, Mathews and Train (2020) adopt the strategy that the best defense is a good offense

[2] I submitted this correction to Ecological Economics in 2016 but the editors told me that I had already "had my say" and refused to consider it for publication.

and inspiring the title of this reply[3] (with apologies to Shakespeare). More seriously, I don't believe that their approach is the best way to advance economic science.[4]

My comment on Desvousges, Mathews and Train (2015) in Ecological Economics addressed three main issues: (1) the data are flawed/low quality, (2) implementation of the adding-up test in the survey is flawed and (3) additional statistical tests for adding-up do not support the Desvousges, Mathews and Train (2015) results. None of these issues are refuted by Desvousges, Mathews and Train (2020). Instead, each of these issues has been confused by the Desvousges, Mathews and Train (2020) "Reply to Whitehead".

In the next section of this I provide some historical context for this latest back and forth between myself and Desvousges, Mathews and Train.[5] In short, following the BP/Deepwater Horizon oil spill, Desvousges, Mathews and Train were funded by BP to discredit the CVM (which can be used to estimate compensable natural resource damages). Then I provide a correction for my mistake and a detailed response to each of the other 11 "mistakes" that Desvousges, Mathews, and Train (2020) claim that I make. It is not clear to me at all that 11 of these are mistakes, per se. Reasonable people can interpret these issues differently. But, from my vantage point, Desvousges, Mathews and Train (2020) have gone over-the-top in their defense of their BP-funded criticism of the CVM.

[3] Wikipedia describes this quote thusly: “The phrase is used in everyday speech to indicate doubt of someone's sincerity, especially regarding the truth of a strong denial.”

[4] See Whitehead (2017) for another unfortunate comment/reply/reply situation.

[5] This section is updated and revised from my November 2016 essay in the AERE Newsletter.

Some History

There was already some bad blood between me and these authors in the context of the validity of the CVM. The story begins in 1989 with the Exxon Valdez oil spill and me working on my dissertation in a graduate student office while watching the news on a 13-inch black and white television. My dissertation, looking at the effects of substitutes on the value of natural resources that generate existence value estimated with the contingent valuation method (CVM) was, unfortunately, timely. I defended in October 1989, a couple of months after I began my first semester as an assistant professor at East Carolina University (ECU). Since I lived only two hours away from the workshop that Kerry Smith ran at North Carolina State University (NCSU) I was fortunate to be exposed to many of the players and much of the economic research surrounding the spill. At the time, however, I was naive about how big of a deal the spill was legally and for the economics profession. For example, I was invited to present the lead paper from my dissertation at the NCSU workshop and was looking for some comments on how Glenn Blomquist (my dissertation advisor) and I might improve the paper. Instead there seemed to be weird pointed questions and a side comment or two about how I should be working for Exxon (my thesis was that existence values may be biased upwards due to lack of information about substitutes). I had absolutely no idea where those comments were coming from because environmental economists didn't work for Exxon (at least they had never advertised in the JOE), did they?

Fast forward. After spending a little over twenty years working on issues surrounding the CVM and other stated preference methods, I was invited to work as a consultant for the State of Florida on a damage assessment resulting from the BP/Deepwater Horizon oil spill (Huffaker, Clouser and Larkin 2012). The purpose of this section of the paper is to present a first person account of the odd experiences associated with that effort as background for the current flap, but I’ll go even further back and describe the first “CVM debate” for some context.

The Exxon Valdez Oil Spill and the CVM Debate

As alluded to above, following the Exxon Valdez oil spill the state of Alaska hired consulting economists to conduct a study estimating economic damages. Exxon hired consulting economists to refute those studies. The argument became known as "the contingent valuation debate." At the time of the spill, the CVM was a promising nonmarket valuation approach with two book length treatments published in the late 1980s. Cummings, Brookshire and Schulze (1986) with funding from the Environmental Protection Agency (EPA), staged a conference that resulted in a book that looked critically at the method with little of the harsh criticism that would come a few years later. The focus was on developing “referencing operating conditions” that would make the CVM more accurate.

As I was writing my dissertation and reading the literature, I became aware of the second book on CVM that was written in the 1980s. Mitchell and Carson (1989) was cited in a number of journal articles in the late 1980s as forthcoming or in press. I was anxious to get a copy as it seemed to make a lot of points that would address many of the problems about the validity and reliability of willingness to pay estimates from the CVM. It still addresses many of the issues that remain contentious in such a way that I wonder what all the fuss is about. Of course, this attitude is naive when millions of dollars are at stake.

Following the Exxon Valdez spill there was much research activity surrounding the

CVM. I attended a number of sessions and panels at the ASSA and SEA meetings. It was exciting for a young economist to hear the big shots discussing things that I was working on. Researchers funded by Exxon were extremely negative about the CVM when some of them had seemed fairly optimistic about the method before the oil spill. Researchers funded by the State of Alaska and, I think, the Federal Government (this one was a top-secret study) were more positive. None of the positive or negative results associated with the NRDA had been published at this point (and there was no such thing as the WWW or PDF) but everyone except me seemed to know what these results looked like.

I went to the ECU library and checked out the Exxon-funded CVM critique book by Hausman et al. (1993) and paid the fines after I kept it way beyond the due date.[6] The transcripts of the Q&A after the Exxon-funded studies were presented at the Exxon-funded conference made the conference sound quite contentious. I attended the conference on the CVM organized by the Department of Energy and the EPA in Reston, Virginia (Bjornstad and Kahn, 1996). This was a fascinating few days as it was especially combative with Exxon-funded, State of Alaska and Federal Government (I think) funded economists on the program and in attendance. Sparks flew. The *Journal of Economic Perspectives* (*JEP*) published a symposium on the CVM in 1994 (Portney 1994, Hanemann 1994, Diamond and Hausman 1994). I always wonder what a naïve reader thinks when picking up these *JEP* articles, but they remain the best summary of the issues and strident tone that arose around the CVM in the early 1990s.

Finally, I attended the 1996 AERE Workshop on combining revealed and stated

[6] I've since bought the book.

preference data (along with my ECU colleagues Tim Haab and Ju-Chin Huang[7]). The theme of this conference seemed to be informally “beyond CVM.” Conjoint analysis (these days, “discrete choice experiments” or DCE) was the new thing and proven by assertion to be far superior to the silly CVM. It seems as though everyone stopped doing CVM and started doing DCE at about this time. I was inspired to pursue more research on joint estimation of revealed and stated preference data and had hopes of moving my own research agenda “beyond CVM.”

BP/Deepwater Horizon (DWH) Blowout and the Revival of the CVM Debate

The BP/DWH blowout occurred on April 20, 2010. A subsequent email sent to the RESECON listserv announced the second edition of the original Exxon-funded Desvousges et al. (1992) Research Triangle Institute (RTI) Nonuse Values study (Whitehead, 2010). This was my first inclination that BP would not be a fan of the CVM and that the CVM debate wasn’t over. I signed a consulting contract with the University of Florida late in 2010, our work began, and we produced a report by March 2012. A settlement was reached in June 2015 and the State of Florida gave us permission to release our results in November 2015.[8]

I still wasn’t totally convinced that the CVM was going to be an issue for the economics profession (hadn’t CVM been replaced by DCE?) until I was asked to read a draft of Kling et al.

[7] Sour grapes / I told you so: the abstract that we submitted to the workshop was rejected but went on to be published in *JEEM* ☺.

[8] See Whitehead et al. (2018) for the first paper from the NRDA study. A second has been accepted for publication (Whitehead, et al. forthcoming).

(2012) and provide comments for a second *JEP* symposium on contingent valuation. Their paper is a very balanced and accurate appraisal of the state of the art of the CVM. There are plenty of warts on the CVM but it is a method that can provide some useful information in a number of contexts. This also was the theme of Richard Carson's (2012) *JEP* article. Jerry Hausman's (2012) piece, provocatively titled "Dubious to Hopeless," came down pretty hard on the CVM.

After the *JEP* symposium appeared, I wrote a blog post pointing out that Hausman's piece did not really review the CVM literature and there had been, in fact, some new developments since 1994 (Whitehead, 2012). One email led to another and Tim Haab (my co-blogger at www.env-econ.net), Matt Interis, Dan Petrolia and I wrote a long cathartic comment without much hope of publication. We didn't even bother to ask the *JEP* if they would entertain the idea of publication. We proposed the paper to the editors of AERE's *Review of Environmental Economics and Policy* but they felt that a more balanced approach was necessary (i.e., a full-blown symposium). Fortunately, the editors of the AAEA's *Applied Economic Perspectives and Policy* were willing to consider the paper. We submitted it, it was sent out for review, revised in response to reviews and published (Haab et al. 2013).

During the time that the *AEPP* paper was in the publication process, I presented the outline during a keynote lunchtime talk at the CNREP 2013 conference (titled "Contingent Valuation: From Dubious to Hopeless?").[9] One of my slides mentioned my constant worry that since MIT vs. App State isn't a fair intellectual fight, there was some hesitation on my part to take on this challenge. In fact, I was terrified that we were making a huge mistake in sticking our necks out. But the only pushback that we received was a comment on our paper by Desvousges,

[9] Click here for Program and Abstracts.

Mathews and Train (2016). In their comment they state that they were unaware of studies that correct for hypothetical bias, focus on the adding up test as a test of the adequacy of scope and reject empirical results from the consequentiality literature, among other things. We wrote a reply with a long and cheeky title (Haab et al. 2016). We provided references to papers that correct for hypothetical bias, describe studies that assess adequate/plausible sensitivity to scope without the adding up test, and admit that the measurement of consequentiality warrants additional study.

Adequate/Plausible Responsiveness to Scope

At this point I'm losing track of the timeline. But sometime in 2011, I think, I agreed to review a paper for *Ecological Economics* that contained a broad but shallow literature review, negative adding-up test evidence from an unpublished NRDA (unpublished beyond obscure court documents), and a negative conclusion about the CVM. I recommended rejection with some suggestions that would make the paper more scholarly and publishable (e.g., a meta-analysis of the determinants of scope effects). When I received a revised version of the paper to review it was much improved but still had some issues. I made some more suggestions and recommended a revise and resubmit. The next time I saw the paper it was published (Desvousges, Mathews and Train 2012). I tried to ignore this development but given that I didn't think the paper had reached its potential I decided to write a comment. My comment turned into a full-blown paper with empirical results from data borrowed from the literature and another cheeky title (Whitehead 2016a). I'll briefly describe these two papers.

Desvousges, Mathews and Train (2012), in the context of an NRDA case unrelated to BP/DWH, review the scope effects literature in search of any paper that demonstrates "adequate" responsiveness.[10] This phrase is from the National Oceanic Atmospheric Administration (NOAA) Panel on Contingent Valuation (Arrow et al. 1993) but as Arrow et al. (1994) point out, a better term is plausible.[11] Nevertheless, Desvousges, Mathews and Train (2012) equate adequate responsiveness with the “adding up test” which was suggested by Hausman (1993). Arrow et al. (1994) also point out that assessing adequacy is about achieving economic, in addition to, statistical significance. Nevertheless, Desvousges, Mathews and Train (2012) summarize their review in terms of statistical significance. Studies that have statistically significant scope effects are classified as studies that “pass” the scope test. Studies that do not have statistically significant scope effects are classified as studies that “fail” the scope test. Studies that include both statistically significant and insignificant scope effects are classified as studies with “mixed” results. They find that only a little more than a third of all scope effect studies “pass.”

A number of my own papers were included in the review so I know intimately that this classification system is a superficial way to summarize the literature. One of my papers was classified as “pass.” The version that we submitted to the journal included a number of sensitivity tests showing under which specifications the scope coefficient was statistically significant or not (this version is available upon request). In one of those journal reviews that you’ll never forget an anonymous referee commented that this sort of sensitivity analysis was “nonsense.” The editor agreed so we presented a single specification that “passed” the test.

[10] See also the comment and reply in *Ecological Economics* (Chapman et al. 2016, Desvousges, Mathews and Train 2016).

[11] Later I realized that Kerry Smith and Laura Taylor pointed out that Arrow et al. clarified the adequate/plausible terminology 20 years ago (Smith and Osborne 1996, p. 288). My apologies to them for the oversight in Whitehead (2016).

Another study was classified as having “mixed” results. In this paper, again, we made an honest effort to show if and when the data exhibited sensitivity to scope. As a result of including this sensitivity analysis the study was classified as “mixed” in the Desvousges, Mathews and Train (2012) classification. I do not doubt that a number of other studies classified as having “mixed” results have authors who might object to the classification.

Desvousges, Mathews and Train (2012, 2016) claim that only a few studies present a test of scope adequacy. In their view the adding up test is the only adequacy test. In doing so, they have overlooked the fact that most every scope effect paper presents information on adequacy/plausibility. Rarely does a scope effect paper present the statistical test without presenting the effect size – the difference in willingness to pay for base and scope conditions. It is this difference that can be used to judge the “plausibility” of scope effects. In Whitehead (2016a), I show that a simple scope elasticity statistic can be used to assess plausibility and provide several examples.

A Revealed Preference Estimate of the Damages of the BP/DWH Oil Spill

In a non-State of Florida NRDA-funded effort, Alvarez et al. (2014) estimated the recreational fishing losses from the oil spill to be $585 million. Kenneth Train called three of the four authors on the phone (I was the lucky one) raising issues that he felt rose to the level of retraction from the *Journal of Environmental Management* (*JEM*) (DuBois 2016). I had been an associate editor at *JEM* for four years and had not known an economics paper to draw such interest. We declined the offer to retract the paper and, instead, I sent the editor of *JEM* an email describing the situation and asking if we could write an addendum with some sensitivity analysis clarifying some of the issues raised in those phone calls. The editor was open to the idea but made it clear that the addendum would be a corrigendum (i.e., correction). While we didn’t think we had made a mistake worse than oversimplification of the aggregation rule, we wrote the corrigendum and it was published (Alvarez et al. 2015a). The corrigendum presented $78 million as our best estimate of the recreational fishing damages from the oil spill.

Kenneth Train wrote a comment on the paper/corrigendum and we wrote a reply with an introduction that contained some of the backstory to provide context. The associate editor sent both the comment and reply out for peer review. We were asked to maybe tone down our response, which we did in a revision, and both were published (Train 2015, Alvarez et al. 2015b). The comment focused on several standard issues in recreation demand such as the appropriate cost per mile. We used a cost per mile estimate that is at the upper end of the range, from the IRS reimbursement rate that includes fixed and variable costs. Train wrote that this is incorrect and no one else has used such an inappropriately high number. Hang, McFadden, Train and Wise (2016) pushed the cost per mile issue even further and conduct a literature review and empirical test of variable vs total costs per mile. Hang et al.’s literature review finds that the most egregious sin is a failure to report the source of the cost per mile estimate. Another objection raised by Train (2015) is “time travel”. The Alvarez et al. (2014) model is intertemporal and when fishing sites close, anglers are allowed to go back in time to fish at other sites. While time travel is dangerous and should never be done without the proper training, its use in our model leads to a more conservative damage assessment (which BP shouldn’t find objectionable). The study was not part of the Florida NRDA, but NRDA is the appropriate context and conservative assumptions are preferred.[12]

The BP-funded McFadden and Train (eds.) book

Daniel McFadden and Kenneth Train edited a book that was published in 2017 that was critical of the CVM (McFadden and Train 2017). I was invited to write a review of this book for the *Australian Journal of Agricultural and Resource Economics* (Whitehead 2018). All but one of the chapters (McFadden's chapter 6) was funded by BP and the primary purpose of the book is to discredit stated preference (both CVM and DCE) approaches for the measurement of non-use values. McFadden and Train's effort is not original. It follows the Exxon-funded strategy of conducting research by hiring economic consultants to examine the ability of the CVM to estimate economic values suitable for NRDA (Maas and Svorenčík, 2017). McFadden and Train (2017) follow this same path in attempting to discredit the CVM (and DCE). A large number of consultants from several major firms were hired, studies were conducted and another book was published critical of the CVM (and DCE). Several chapters focus on known weaknesses of stated preference methods and some focus on newer issues. Some of the data collected with funding from BP is of relatively low quality (e.g., Desvousges, Mathews and Train (2015) was reprinted as Chapter 4). Overly broad generalizations are made from these studies without placing them in the context of the literature. The studies are conducted by researchers with limited experience in stated preference methods. The data analysis for each chapter leaves many questions.

[12] It did not escape my attention that papers that I authored or co-authored are disproportionately cited by Hang et al. (2016). This might lead a naïve reader to think that we are two of the leading lights in the recreation demand literature. I assure the reader that, at least in my case, I have not risen above practitioner status.

Questionable assumptions are made (e.g., the cutoff for plausible scope elasticity is 0.20). A good objective round of peer-reviews would have tightened each chapter.[13]

My "Mistakes" in Whitehead (2020)

Before I describe my mistakes, I'll review the dichotomous choice contingent valuation, data analysis and results (see Whitehead 2017). Survey respondents are presented with a randomly assigned dollar amount which is the purported individual or household cost of a government policy. Respondents indicate whether they are willing to pay for this policy. These data are typically analyzed with a logit or probit model where the 0,1 indicators of willingness to pay are regressed on the dollar amount. The bid curve is plotted with the proportions of respondents who are willing to pay on the vertical axis and the cost amount on the horizontal axis. With idealized data, such as in Figure 1, the percentage of respondents willing to pay is decreasing with the cost amount, 100% of respondents would be in favor of the policy if the cost amount was $0 and close to 0% would pay at the highest cost amount. In this situation it does not matter how the data are analyzed for estimation of willingness to pay (Landry and Whitehead 2020). Real-world dichotomous choice data rarely meets this ideal.

In Whitehead (2020) I describe the Desvousges, Mathews and Train (2015) dichotomous choice data as full of non-monotonicities, flat portions of bid curves and fat tails. A non-monotonicity is when the percentage of respondents in favor of a policy increases when the cost increases. In other words, for a pair of cost amounts it appears that respondents are irrational when responding to the survey. This problem could be due to a number of things besides

[13] See Navrud (2018) for a similar description of limitations about the book.

irrationality. First, respondents may not be paying close attention to the cost amounts. Second, the sample sizes may be simply too small to detect a difference in the correct direction. Whatever the cause, non-monotonicities increase the standard errors of the slope coefficient in a parametric model.

Flat portions of the bid curve exist when the estimated bid curve may be downward sloping but the slope is not statistically different from zero. This could be caused by small differences in cost amounts and/or it is due to sample sizes that are too small to detect a statistically significant difference. For example, there may be little perceived difference by survey respondents between a cost amount of $5 and $10 relative to a larger difference, say $5 to $50. And, even if the percentage of responses in favor of a policy is economically different between two cost amounts, this difference may not be statistically different due to small sample sizes.

Fat tails may exist when the percentage of respondents who are in favor of a policy is high at the highest cost amount. However, this is only a necessary condition. A sufficient condition for a fat tail is when the percentage of respondents who are in favor of a policy is high at two or more of the highest cost amounts. In this case, the fat tail will cause a parametric model to predict a very high cost amount that drives the probability that respondents are in favor of a policy to (near) zero. A fat tail will bias a willingness to pay estimate upwards because much of the WTP estimate is derived from the portion of the bid curve when the cost amount is higher than the highest cost amount presented in the survey.

Desvousges, Mathews and Train (2015) use the contingent valuation method with the nonparametric Turnbull estimator and find that the data do not pass the adding-up test. In

Whitehead (2020) I address three issues with Devousges, Mathews and Train (2015): (1) the data are of low quality, (2) implementation of the adding-up test in the survey is flawed and (3) additional statistical tests for adding-up do not support the authors' adding-up conclusion. Desvousges, Mathews and Train (2020) reply to my comment by describing 12 mistakes.

Mistake #3: I botched the t-tests

I agree that I made one of the mistakes on the list. I conducted an adding-up test by examining whether the confidence intervals for two willingness to pay (WTP) estimates (the whole and the sum of the WTP parts) overlap. It is well-known that confidence intervals can overlap and yet the t-statistic for the test will indicate that the difference in means is statistically different. The mistake that I made was not checking the t-statistic.

In addition (and not pointed out by Desvousges, Mathews and Train 2020), I used inappropriate confidence intervals for this test. In Whitehead (2020) I used the symmetric confidence intervals produced by the Delta Method, a first-order Taylor Series expansion from the variance-covariance matrix (Cameron 1991). However, 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 that is common in the contingent valuation literature is the Krinsky-Robb confidence interval (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, along with Fieller and bootstrap approaches, and finds little difference for well-behaved (simulated and real) data (such as is imagined in Figure 1). However, Hole (2007) points out that WTP must be normally distributed for the Delta Method confidence interval to be accurate. I used NLogit (www.limdep.com) software to estimate the confidence intervals in Whitehead (2020). NLogit allows for both the Delta Method and Krinsky-Robb approaches. But, the Krinsky-Robb confidence interval estimation approach imposes symmetry. Hole (2007, p. 831) explains how: "The [Krinsky-Robb] confidence interval could also be derived by using the draws to calculate the variance of WTP ..., but this approach, like the delta method confidence interval, hinges on the assumption that WTP is symmetrically distributed." When estimating Krinsky-Robb confidence intervals in NLogit, four of the five "blew up" when using the Desvousges, Mathews and Train (2015) data. The upper and lower limits were in the +/- 10s and 100s of thousands, respectively. In my experience, Delta Method and Krinsky-Robb confidence intervals are not much different when estimated in NLogit when the data are relatively well-behaved (as shown below).

The logit models for the Desvousges, Mathews and Train (2015) and Chapman et al. (2009) whole scenario data are presented in Table 1.[14] In each model the constant and the coefficient on the cost amount are statistically different from zero. But, the precision of the cost coefficient with the Desvousges, Mathews and Train (2015) data is low relative to that from the Chapman et al. (2009) data. The WTP estimates from the linear model and confidence intervals are presented in Table 1. The Delta Method confidence intervals are estimated in NLogit and the Krinsky-Robb confidence intervals are estimated in SAS (www.sas.com) without imposing

[14] The Desvousges, Mathews and Train (2015) data are collected with a survey based on the Chapman et al. (2009) study. More detail is provided in Whitehead (2020).

symmetry.[15] The Delta Method lower bound on the Desvousges, Mathews and Train (2015) WTP estimate is 44% of the Krinsky-Robb lower bound. The Krinsky-Robb upper bound is 269% larger than the Delta Method upper bound. The imprecision of the coefficient on the cost amount is driving the asymmetry revealed by the Krinsky-Robb approach. These results should be considered in contrast to the WTP estimate from the Chapman et al. (2009) data (the study that Desvousges, Mathews and Train (2012, 2015) criticize for failing the adding-up test). With these data the Delta Method and Krinsky-Robb intervals are symmetric and similar due to the precision of the cost amount estimate.[16]

The point estimate of the sum of the WTP parts is $1114 estimated from similar models with the Desvousges, Mathews and Train (2015) data (Whitehead 2020). The WTP for the sum of the parts is within the Krinsky-Robb interval for the whole scenario [$234, $2062] indicating that we cannot reject the hypothesis that WTP for the whole is equal to WTP for the sum of the parts (p = 0.05). These results are consistent with my 2020 conclusion that the willingness to pay estimates in Desvousges, Mathews and Train (2015) are not estimated precisely enough to conclude that these data fail the adding-up test (Whitehead 2020).

[15] The Krinsky-Robb intervals are akin to what Hole (2007) calls the Monte Carlo percentile approach. I take one million draws from the variance-covariance matrix and trim the $\alpha/2$ highest and lowest WTP values, where $\alpha=0.05$. Hole's (2007) Krinsky-Robb intervals are based on a resampling approach and the Monte Carlo approach.

[16] The Krinsky-Robb confidence interval estimated in NLogit with the Chapman et al. (2009) data is [237, 320] which is also very similar to the Delta Method interval.

Mistake #1: The log-linear models are meaningless

When dichotomous choice CVM data has a negative WTP problem (see mistake #2), one of the standard corrections is to estimate a log-linear model and present the median WTP. With many estimated log-linear models the mean WTP is undefined. This is because the log-linear model flattens the estimated survival curve and, in contrast to a linear model, the probability of a no response does not approach zero at any reasonable bid amount. The median WTP estimate from the log-linear model tends to be a useful supplement to the welfare measures available from the linear model.

Desvousges, Mathews and Train (2020) rightly argue that (1) the sum of medians is not equal to the median of the sums and (2) the mean WTP estimates for each of their five scenarios are basically infinity (or in the millions of dollars when the data are estimated with a log-linear probit). Given this empirical fact, there seems to be no median estimate available for the sum of the four individual WTP amounts (that could be compared to the median of the whole). However, Desvousges, Mathews and Train (2020) are able to estimate the "median of the sum of WTPs through simulation" to be $4904. They then explain that since $4904 is 24 times that of the median of the whole scenario, $201, it is "clearly a violation of adding up". Given that the confidence interval from the Delta Method, [-47, 449], does not include $4904 we could reject the notion that the willingness to pay estimates pass this version of the adding-up test.

However, above I argue that standard errors from the Delta Method are likely the wrong ones to use since the cost parameter is measured without much precision. In this case the Krinsky-Robb confidence intervals are more appropriate. The Krinsky-Robb confidence interval for the median WTP estimate for the whole scenario is [54, 9558]. Since the median of the sum of the WTP estimates from the four adding-up scenarios, estimated by Desvousges, Mathews and Train (2020) to be $4904, lies within the 95% Krinsky-Robb confidence interval then we fail to reject the adding-up hypothesis at the 95% confidence level.

The only other adding-up test that can be conducted with median WTP estimates is to compare the median for the whole with the sum of the medians for the four parts. In this test I found that the median WTP estimates passed the adding up test using the confidence intervals from the Delta Method (Whitehead 2020). This test is a useful supplement when one is inclined to consider the robustness of the adding-up test conducted with only the Turnbull estimator (as in Desvousges, Mathews and Train 2015). Otherwise, researchers are treating mean WTP estimates in the millions as if they are meaningful.

Mistake #2: The use of models that allow negative willingness to pay is inappropriate

When dichotomous choice CVM data is of low quality, the measure of central tendancy is sensitive to assumptions. In contrast, with the highest quality data it makes no difference which WTP estimator that is used (Landry and Whitehead 2020). As data quality falls, however, the choice of WTP estimate can matter a great deal. In this situation, so as to avoid sponsor and other biases, it is important for the CVM researcher to present the full range of WTP estimates and avoid the impression that results have been cherry picked. This range of WTP provides a more complete depiction of analyst uncertainty and allows for sensitivity and other analyses.

I have grown suspicious over the years whenever I see hypothesis tests conducted with only the Turnbull WTP estimate such as in Desvousges, Mathews and Train (2015). First, the Turnbull is a lower bound WTP estimate and potential differences across treatments are minimized. Second, its standard errors are smaller than parametric WTP estimates estimated with the same data. This second observation is due to the way that the standard errors are calculated and to the fact that the data are “smoothed” (i.e., the dependent variable is recoded) when there are non-monotonicities. As Haab and McConnell (1997, p 253) explained (emphasis added): "We demonstrate that the Turnbull model ... provides a straightforward alternative to parametric models, **so long as one simply wants to estimate mean willingness to pay**." In my opinion, the Turnbull WTP estimate should not be used for hypothesis testing. When hypothesis tests are being conducted, a range of WTP estimates should be used to determine if the results are robust to estimation method.

So, is it reasonable to include the linear-in-bid parametric model in the collection of WTP estimates in Whitehead (2020)? Hanemann (1984, 1989) showed that in a linear utility model the mean (and median) willingness to pay is WTP = $-\alpha/\beta$, where α is the constant in a logit or probit model and β is the coefficient on the randomly assigned cost amount. One benefit of this estimate is that it is insensitive to fat tails. However, this estimate allows for negative WTP values unless the probability of a yes response to a dichotomous choice question is 100% when the bid amount is zero. Negative WTP values can enter into the analysis in two ways. First, the WTP estimate itself can be negative. This will occur when the probability of a yes response at the lowest bid amount is less than 50%. The second possibility is that the empirical distribution of WTP can include negative values. This is of little consequence to the analysis unless the confidence interval includes zero. Both circumstances arise with the Desvousges, Mathews and Train (2015) data.

Desvousges, Mathews and Train (2020) dismiss outright the possibility of negative WTP, rendering the linear model inapppropriate. Their dismissal is consistent with Haab and McConnell's argument that since public goods are freely disposable, negative WTP is only an empirical artifact of a distributional assumption. But, with government policy free disposal is not always possible. In the case of a cleanup of natural resource damages, the cleanup could be considered a wasteful intrusion into a private business decision.[17] Considering this, I would not be surprised if some of the respondents to CVM scenarios demanded compensation for environmental cleanup.

There have been a number of suggestions about how to handle negative willingness to pay. Many of these involve obtaining more data with follow-up questions (Landry and Whitehead 2020). Unfortunately, the Desvousges, Mathews and Train (2015) survey data does not have any of this supplemental information. In that case, an assumption that negative WTP is a possibility cannot be ruled out. Inclusion of the linear model allowing for negative WTP, as long as it is presented along with other estimates, should not be dismissed outright.

Desvousges, Mathews and Train (2020) state: "This means that adding-up passed in his calculations on linear models not because of the data but because of his implausible additional assumption that many people have a negative WTP for the environmental programs." It is not true that the linear model finds that "many" people have negative willingness to pay in each of the scenarios. According to my Krinsky-Robb WTP simulation with the Desvousges, Mathews and Train (2015) data, the percentage of negative WTP values for the whole, first, second, third,

[17] Bohara, Kerkvliet and Berrens (2001) discuss how and why negative WTP values might arise, along with empirical examples.

and fourth scenarios in the Desvousges, Mathews and Train (2020) data are 2%, 0.01%, 77%, 25% and 0.83% for the first through fifth scenarios. The WTP from the second scenario is negative due to less than 50% being willing to pay the lowest cost amount. The WTP from the third scenario has a Delta method confidence interval that includes zero.

If the negative mean WTP from the second scenario is set equal to zero then the adding-up test is still supported. The Krinsky-Robb confidence interval is [68, 788] which includes the sum of the WTP for the parts with WTP from the second scenario set equal to zero ($467) indicating that the adding-up test is supported.

Mistake #4: I misused the weighted model results

Desvousges, Mathews and Train (2020) drew attention to my treatment of the weighted WTP estimates. The regression model for the second scenario has a negative sign for the constant and a positive sign for the slope. Desvousges, Mathews and Train (2020)complain that I "mechanically" calculate WTP for the second scenario and this positive number adds weight to the sum of the WTP parts. This is in contrast to the unweighted data for which WTP is negative. They argue that inclusion of the data from this scenario biases the adding-up tests in favor of the conclusion that the WTP data does not pass the adding-up test. Desvousges, Mathews and Train's (2020) complaint has merit but it distracts from the true problem.

The motivation for my consideration of the weighted data was Desvousges, Mathews and Train's (2015) claim that they found similar results with the weighted data. My analysis with the weighted data uncovered validity problems with two of the five scenarios which, when included in an adding-up test, led to a failure to reject adding-up. At this point it is instructive to visually examine Desvousges, Mathews and Train's (2015) weighted data to see if it even passes the "laugh" test.

In Figure 2 the weighted willingness to pay responses and the Turnbull curve for the whole scenario are displayed.[18] The dots and dotted lines represent the raw data. Instead of a downward slope, these data are "roller-coaster" shaped (two scary hills with a smooth ride home). The linear probability model (with weighted data) has a constant equal to 0.54 (t=9.73) and a slope equal to -0.00017 (t=-0.69). This suggests that the whole scenario data, once weighted, lacks validity because the slope coefficient is not statistically different from zero.

While lacking validity, the solid line Turnbull illustrates how researchers such as Desvousges, Mathews and Train (2015) can obtain a WTP estimate with data that does not conform to rational choice theory. The Turnbull smooths the data over the invalid stretches of the bid curve (the "non-monotinicities") and the WTP estimate is the sum of the area of the rectangles. In this case WTP = $191 which is very close to the unweighted Turnbull estimate. But, a researcher should consider this estimate questionable since the underlying data does not conform to theory. As a reminder, the WTP for the whole scenario is key to the adding up test as it is compared to the sum of the parts. The WTP estimate from the weighted linear logit model is $239 and the Krinsky-Robb [-8938, 9615] confidence interval includes zero. Given the statistical uncertainty of the WTP estimate, it is impossible to conduct any sort of hypothesis test with these weighted data.

[18] Note that the weights are scaled to equal to sub-sample sizes.

The weighted votes and the (pooled) Turnbull for the second scenario, the scenario that drew Desvousges, Mathews and Train's complaint, are presented in Figure 3. Instead of a downward slope, these data are "Nike swoosh" shaped. The linear probability model (with weighted data) has a constant equal to 0.13 (t=2.46) and a slope equal to 0.00107 (t=4.19). This suggests that the second scenario data, once weighted, lacks validity because the slope coefficient has the wrong sign. Again, the Turnbull estimator masks the weakness of the underlying data. In this case, the Turnbull is essentially a single rectangle. With pooling the probability of a vote in favor is equal to 28.06% for the lower bid amounts. With pooling the probability is 27.56% for the higher bids. The Turnbull WTP estimate is $112 which appears to be a reasonable number, hiding the problems with the underlying data.

Desvousges, Mathews and Train (2015) were technically correct when they asserted that they found similar results with the weighted data. Their assertion is true when the Turnbull WTP estimates are developed with smoothed data. But, the underlying weighted data is a mess that is hidden by the Turnbull smoothing.

Desvousges, Mathews and Train (2015) re-estimated the full data model with the cost coefficients across whole and treatments constrained to be equal. In a utility difference model the cost coefficient is the estimate for the marginal utility of income. There is no reason for marginal utility of income to vary across treatments unless the clean-up scenarios and income are substitutes or complements. This theoretical understanding does not explain why the weighted models for the whole and second scenarios are not internally valid (i.e., the cost coefficient is not negative and statistically different from zero). The model that Desvousges, Mathews and Train refer to passes a statistical test, i.e., the model that constrains the cost coefficients to be equal is not worse statistically than an unconstrained model, but it should be considered inappropriate due to the lack of validity in the weighted whole and second scenario data sets. Desvousges, Mathews and Train's (2015) use of the model with a constrained cost coefficient amounts to hiding a poor result. The reason that the weighted model with the full data set takes the correct sign is because the scenarios with correct signs outweigh the scenarios with incorrect or statistically insignificant signs. The reader should attach little import to Desvousges Mathew and Train's (2015) claim that their result is robust to the use of sample weights.

Mistake #5. I didn't present the adding up test using the Kriström estimator

Desvousges, Mathews and Train (2020) notice that I conducted an adding-up test with the Kriström nonparametric estimator in a 2016 blog post.[19] They claim that I "inadvertently dropped observations" when conducting these calculations. Dropping these observations was not "inadvertant." In the blog post I used a sample size of n=950 which is the same sample size that Desvousges, Mathews and Train (2015) used in their Table 5 (dropping a handful of observations with a missing age variable).

Further, Desvousges, Mathews and Train (2020) report that the adding-up test fails with the Kriström estimator and I "failed to report relevant findings" because I did not include this in Whitehead (2020). I dropped the Kriström estimator WTP estimates because it requires high quality data that can be used to estimate a reliable choke point for the referendum vote demand

[19] https://www.env-econ.net/2016/09/comment-on-desvousges-mathews-and-train-2015-continued-wtp-estimates-.html

curve. With low quality data, as in Desvousges, Mathews and Train (2015), any choke price estimate is too ad-hoc for conducting statistical tests.

The second complaint begs the question: how many additional tests should be conducted in a comment on a paper? In Whitehead (2020) I provided three parametric tests using some the standard models in the literature. I then consider the robustness of these tests with (a) weighted data and (b) the complete case data set (n=934 after dropping those with missing age and income). Most of these tests do not support failure of the adding-up hypothesis.

Mistake #6. My data in many of my publications is just as bad as the data in Desvousges, Mathews and Train (2015)

Desvousges, Mathews and Train (2020) state that these data problems also occur in Chapman et al. (2009) and a number of my own CVM data sets. They are correct. But, Desvousges, Mathews and Train (2020) are confusing the existence of the problem, in the case of non-monotonicity and flat portions, with the magnitude of the problem. And, they are assuming that if the necessary condition for fat tails exists then the sufficient condition also exists. Many, if not most, CVM data sets will exhibit non-monotonicities and flat portions of the bid curve. But, these issues are not necessarily an empirical problem.

To illustrate that the Desvousges, Mathews and Train (2015) data are actually worse than my data I estimated the logit model, WTP value and 95% Krinsky-Robb confidence intervals for 20 data sets. Five of the data sets are from Desvousges, Mathews and Train (2015), 2 are from Chapman et al. (2009) and 13 are from my papers published between 1992 and 2009 referenced by Desvousges, Mathews and Train (2020).[20] The average sample size for these 20 data sets is 336 and the average number of randomly assigned cost amounts offered to respondents is 5.45. The average sample size per cost amount is 64, which is typically sufficient to avoid data quality problems. A good rule of thumb is that the number of data points for each cost amount should be $n \geq 40$ in the most poorly funded study.

These averages obscure differences across study authors. The average sample size for the Desvousges, Mathews and Train (2015) data sets is 196. With 6 cost amounts the average sample size per cost amount is 33, lower than my rule of thumb. The Chapman et al. (2009) study is the best funded study in this sample and the two sample sizes are 1093 and 544. With 6 cost amounts the sample sizes per cost amount are 182 and 91. The Whitehead studies have an average sample size of 317 and with an average of 5 cost amounts, the sample size per cost amount is 65. Already, differences across these three groups of studies emerge with Desvousges, Mathews and Train (2015) study the most poorly designed.[21]

There are a number of dimensions over which to compare the logit models in these studies. My preferred measure is the ratio of the upper limit of the 95% Krinsky-Robb confidence interval for WTP to the median WTP estimate. This ratio will be larger the more extensive are the three empirical problems mentioned above. As this problem worsens,

[20] Desvousges, Mathews and Train (2020) mention 15 data sets but two of the studies use the same data as in another paper.

[21] It is not clear why Desvousges, Mathews and Train (2015) did not conduct a three sample adding-up test as described by Diamond (1996) and Desvousges, Mathews and Train (2012). If they did so the sub-sample sizes may have been sufficiently large.

hypothesis testing with the WTP estimates (again, a function of the ratio of coefficients) becomes less feasible. It is very difficult to find differences in WTP estimates when the confidence intervals are very wide. To suggest that this measure has some validity, the correlation between the ratio and the p-value on the slope coefficient is r = 0.96.

The results of this analysis are shown in Figure 4. The ratio of the upper limit of the confidence interval to the median is sorted from lowest to highest. The Desvousges, Mathews and Train (2015) values are displayed as orange squares, the Chapman et al. (2009) values are displayed as green diamonds and the Whitehead results are displayed as blue circles and one blue triangle. The blue triangle is literally "off the chart" bad so I have divided the ratio by 2. This observation, one of three data sets from Whitehead and Cherry (2007), does not have a statistically significant slope coefficient.

Considering the Desvousges, Mathews and Train data, observation 19, with a ratio of 4.82 (i.e., the upper limit of the Krinsky-Robb confidence interval is about 5 times greater than the median WTP estimate), is the worst data set. Observation 8, the best Desvousges, Mathews and Train (2015) data set, has their largest sample size of n=293. The Chapman et al. (2009) data sets are two of the three best in terms of quality. The Whitehead data sets range from high to low in terms of quality. Overall, four of the five Desvousges, Mathews and Train (2015) data sets are in the lower quality half of the sample.

Of course, data quality should also be assessed by the purpose of the study. About half of the Whitehead studies received external funding. The primary purpose of these studies was to develop a benefit estimate useful for policy analysis. The other studies were funded internally with a primary purpose of testing low stakes hypotheses. In hindsight and years of peer-reviewed literature, these internally funded studies were poorly designed with sample sizes per bid amount too small and/or poorly chosen bid amounts. With the mail surveys in the Whitehead studies the number of bid amounts was chosen with optimistic response rates in mind. In another study a bid amount lower than $100 should have been included. Many of the cost amounts are too close together to obtain much useful information. Hindsight is 20/20.

Mistake #7. I ignored the fact that the Desvousges, Mathews and Train (2015) conveys information about substitution effects to survey respondents.

In Whitehead (2020) I argue that the Desvousges, Mathews and Train's (2015) survey design does not allow for substitution effects. This is based on my review of the Chapman et al. (2009) survey instrument which develops a scope test, not an adding-up test. Desvousges, Mathews and Train (2015) assert that their survey does conveys information about substitution effects to survey respondents. But, Desvousges, Mathews and Train (2015) have not provided their internet survey for my review. I asked twice. The first time Bill Desvousges had his assistant send me the Chapman et al. (2009) report containing their in-person surveys. Desvousges, Mathews and Train (2015) state that they implemented their adding-up tests with minor modifications to the Chapman et al. (2009) survey. The second time I asked to review the survey instrument Bill Desvousges complained to the *Economics: E-Journal* editor about my request. The editor told me that he thought I had everything I needed to write my replication paper and not to email Bill Desvousges again.[22] Claims that their survey conveys information about substitution effects to survey respondents are simply assertions. It would be forthright, and

[22] I won't.

standard in the CVM literature, to provide the survey for review.

Mistake #8. I used the wrong word to describe Desvousges, Mathews and Train's income effect tests.

Desvousges, Mathews and Train (2015) simulate the income effect required in an adding-up test (Diamond 1996). Desvousges, Mathews and Train (2020) are correct by pointing out that "implicit claim" may be poor word choice. Desvousges, Mathews and Train (2015) have an empirical finding that income effects are small when simulating income effects with an empirical model that has a statistically insignificant income coefficient. My "implicit" word choice is related to my comment that an adding-up test requires explicit incorporation into the survey instrument (Diamond 1996). Desvousges, Mathews and Train's (2015) income effect test is required because they (apparently) did not develop a survey instrument to allow an adding-up test.

Note that in mistake #9 Desvousges, Mathews and Train (2020) acknowledge that there is a statistically significant income coefficient when the weighted data are used. They have not explained why they chose to simulate income constraints in their robustness tests instead of incorporating income effects explicitly in the survey design. Again, Desvousges, Mathews and Train have refused to share their survey instrument so that I can review how they adapted the Chapman et al. (2009) survey for their adding-up test.

Mistake #9. My treatment of the income effect is incorrect.

I find that the Desvousges, Mathews and Train's (2015) data models has a positive and statistically significant income coefficient with weighted data. Desvousges, Mathews and Train

(2020) state that they re-ran their simulations with the weighted income coefficient and found similar results. But, if they re-ran their simulations with the weighted income coefficient they should have done the adding-up test with the weighted WTP models. Two of the five weighted model willingness to pay estimates lack validity (see Mistake #4 above).

In Whitehead (2020) I also doubt that income is the correct budget constraint. I suspect that survey respondents have some environmental contribution budget in mind when answering CVM questions. Desvousges, Mathews and Train (2020) state that this is a violation of microeconomic theory (their footnote 4). I assume that they are referring to neoclassical microeconomics and not behavioral economics. Even in the confines of neoclassical microeconomics, a budget constraint is consistent with two-stage budgeting where a household first allocates income to different budget categories and then maximizes subutility functions subject to the budget constraint (Deaton and Muellbauer 1980a). Two-stage budgeting theory led to the development of the Almost Ideal Demand System (AIDS) econometric model (Deaton and Muellbauer 1980b).

Mistake #10. I used a one-tailed adding-test statistical test.

My proposed hypothesis for Desvousges, Mathews and Train's data, based on my read of Desvousges, Mathews and Train (2015) and their refusal to share their survey instrument (see Mistake #7), is a one-tailed scope test. It is not a one-tailed adding-up test.

Mistake # 11. The size of the scope effects are not "adequate".

As I pointed out in the Desvousges, Mathews and Train (2012) inspired Whitehead (2016), Arrow et al. (1994) regret using the term adequate (in reference to the size of scope effects) in the NOAA Panel report. Instead they suggested a more appropriate word is plausible scope effects. I proposed scope elasticity as a plausibility measure in Whitehead (2016). Scope elasticity is a more useful measure of plausibility than the adding-up test is for adequacy given difficulties in conducting an adding-up test.

Mistake #12. My Turnbull standard error estimates differ from Desvousges, Mathews and Train's (2015) standard errors.

I applied the formulas in Haab and McConnell (2002) with pooled (smoothed) data. Desvousges, Mathews and Train (2020) report that they used the raw data to construct confidence intervals with the smoothed data WTP estimate. My estimates of the standard errors are larger than Desvousges, Mathews and Train’s (2020). But, standard errors constructed with the raw data should be larger than standard errors from smoothed data. Desvousges, Mathews and Train (2020) do not provide much information on the estimation of their standard errors so it is difficult to say more.

Conclusion

The effort to criticize a valuation methodology, undertaken on behalf of Exxon to avoid paying damages to the State of Alaska, led to what has become known as the “CVM Debate” (Banzhaf 2017). Similarly, considering the history of the CVM debate and the study's funding source (Ioannidis and Doucouliagos 2013, Maas and Svorenčík 2017), the likely purpose of the Desvousges, Mathews and Train (2015) study is to discredit the CVM in the context of natural resource damage assessment. Randall (1993, 1998) points out that much of the first contingent valuation debate was a matter of looking for the “critical test” that would either reject or fail to reject the method itself. In the 1990s the critical test became the split-sample (external) scope test (Whitehead 2016). The adding up test is the most recent example of the search for the critical test of the CVM (Desvousges, Mathews and Train 2015).

In this context I wrote my 2020 comment on Desvousges, Mathews and Train (2015). Given the funding of the study and its purpose, I should not have been surprised by the tone of their reply. While I argue with Desvousges, Mathews and Train over the minutiae of these tests and different estimators, the reader shouldn't lose sight of how silly the debate over the Desvousges, Mathews and Train (2015) article has become. The bottom line is that the Desvousges, Mathews and Train (2015) data are low quality and do not rise to the threshold that is needed to support any hypothesis test much less an adding-up test, which may be the most rigorous CVM validity test ever proposed. The extent of the three problems in Desvousges, Mathews and Train (2015) is severe -- so severe that it makes their attempt to conduct an adding up test (or any test) near impossible.

Pursuit of the critical test is not the most appropriate way to pursue social science research. Hopefully, future research with stated preference methods will continue much the way it proceeded between 1996 and 2010 (without the high stakes) and the cumulative effect of a large number of studies will lead to a better understanding of the accuracy of the CVM. Research should be conducted to advance knowledge and not in the context of maximizing consulting income and divvying up millions of dollars in court proceedings.

References

Alvarez, Sergio, Sherry L. Larkin, John C. Whitehead, and Tim Haab. "A revealed preference approach to valuing non-market recreational fishing losses from the Deepwater Horizon oil spill." Journal of Environmental Management 145 (2014): 199-209.

Alvarez, Sergio, Sherry L. Larkin, John C. Whitehead, and Tim Haab. "Corrigendum: A revealed preference approach to valuing non-market recreational fishing losses from the Deepwater Horizon spill (Journal of Environmental Management 145, 2014, 199–209)." Journal of Environmental Management 150 (2015a): 516-518.

Alvarez, Sergio, Sherry L. Larkin, John C. Whitehead, and Tim Haab. "Reply to "Comment on: A revealed preference approach to valuing non-market recreational fishing losses from the deepwater horizon oil spill and its corrigendum"." Journal of Environmental Management (2015b).

Arrow, Kenneth, Robert Solow, Paul Portney, Edward E. Leamer, Roy Radner and Howard Schuman. Report of the NOAA Panel on Contingent Valuation. January 11, 1993.

Arrow, Kenneth, Edward E. Leamer, Howard Schuman and Robert Solow, Appendix D in "Comments on Proposed NOAA/DOI Regulations on Natural Resource Damage Assessment." U.S. Environmental Protection Agency. October 1994.

Bjornstad, David J., and James R. Kahn. The Contingent Valuation of Environmental Rresources: Methodological Issues and ResearchNneeds. Edward Elgar Publishing Ltd, 1996.

Bohara, Alok K., Joe Kerkvliet, and Robert P. Berrens. "Addressing negative willingness to pay in dichotomous choice contingent valuation." Environmental and Resource Economics 20, no. 3 (2001): 173-195.

Cameron, Trudy Ann. "Interval estimates of non-market resource values from referendum contingent valuation surveys." Land Economics 67, no. 4 (1991): 413-421.

Carson, Richard T. "Contingent valuation: A practical alternative when prices aren't available." Journal of Economic Perspectives 26, no. 4 (2012): 27-42.

Chapman, David, Richard Bishop, Michael Hanemann, Barbara Kanninen, Jon Krosnick, Edward Morey and Roger Tourangeau. 2009. Natural Resource Damages Associated with Aesthetic and Ecosystem Injuries to Oklahoma's Illinois River System and Tenkiller Lake.

Chapman, David J., Richard C. Bishop, W. Michael Hanemann, Barbara J. Kanninen, Jon A. Krosnick, Edward R. Morey, and Roger Tourangeau. "On the adequacy of scope test results: Comments on Desvousges, Mathews, and Train." Ecological Economics 130 (2016): 356-360.

Cummings, Ronald G., David S. Brookshire, and William Schulze. Valuing Environmental Goods: An Assessment of the Contingent Valuation Method. Rowman & Littlefield Pub Incorporated, 1986.

Deaton, Angus, and John Muellbauer. Economics and Consumer Behavior. Cambridge University Press, 1980a.

Deaton, Angus, and John Muellbauer. "An almost ideal demand system." American Economic Review 70, no. 3 (1980b): 312-326.

Desvousges, Willam H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson and K. Nicole Wilson. Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy. Monograph prepared for Exxon Company, U.S.A., Research Triangle Park: Research Triangle Institute. 1992.

Desvousges, Willam, F. Reed Johnson, Richard Dunford, Kevin Boyle, Sara Hudson and K. Nicole Wilson. Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy. Second Edition, RTI Press, RTI International. (2010).

Desvousges, William, Kristy Mathews, and Kenneth Train. "Adequate responsiveness to scope in contingent valuation." Ecological Economics 84 (2012): 121-128.

Desvousges, William, Kristy Mathews, and Kenneth Train. “From Curious to Pragmatically Curious: Comment on ‘From Hopeless to Curious? Thoughts on Hausman's Dubious to Hopeless Critique of Contingent Valuation”.” Applied Economic Perspectives and Policy 38, no. 1 (2016): 174-182.

Desvousges, William, Kristy Mathews, and Kenneth Train. “Reply to ‘On the adequacy of scope test results: Comments on Desvousges, Mathews, and Train’,” Ecological Economics 130 (2016): 361–362.

Desvousges, William, Kristy Mathews, and Kenneth Train. "An Adding-up Test on Contingent Valuations of River and Lake Quality." Land Economics 91, no. 3 (2015): 556-571.

Desvousges, William H., Kristy E. Mathews, Kenneth E. Train, “Reply to Whitehead,” Ecological Economics, 2020.

Diamond, Peter. "Testing the internal consistency of contingent valuation surveys." Journal of Environmental Economics and Management 30, no. 3 (1996): 337-347.

Diamond, Peter A., and Jerry A. Hausman. "Contingent valuation: Is some number better than no number?" Journal of Economic Perspectives 8, no. 4 (1994): 45-64.

DuBois, Shelley, “How much do oil spills cost? Controversy over paper oozes into larger debate,” Retraction Watch, March 31, 2016.

Landry, Craig, and John Whitehead, "Estimating Willingness to Pay with Referendum Follow-up Multiple-Bounded Payment Cards," paper presented at the 2020 W-4133, Athens, GA, February.

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., Matthew G. Interis, Daniel R. Petrolia, and John C. Whitehead. "From hopeless to curious? Thoughts on Hausman's “dubious to hopeless” critique of contingent valuation." Applied Economic Perspectives and Policy (2013): 35(4):593-612.

Haab, Timothy C., Matthew G. Interis, Daniel R. Petrolia, and John C. Whitehead. "Interesting Questions Worthy of Further Study: Our Reply to Desvousges, Mathews, and Train's (2015) Comment on Our Thoughts (2013) on Hausman's (2012) Update of Diamond and Hausman's (1994) Critique of Contingent Valuation." Applied Economic Perspectives and Policy 38, no. 1 (2016): 183-189.

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, W. Michael. "Valuing the environment through contingent valuation." The Journal of Economic Perspectives 8, no. 4 (1994): 19-43.

Hang, Derrick, Daniel McFadden, Kenneth Train, and Ken Wise. "Is Vehicle Depreciation a Component of Marginal Travel Cost? A Literature Review and Empirical Analysis." Journal of Transport Economics and Policy 50(2):1-19, 2016.

Hausman, Jerry. (ed). Contingent Valuation: A Critical Assessment. Elsevier Science Publishers, Amersterdam, The Netherlands 1993.

Hausman, Jerry. "Contingent valuation: from dubious to hopeless." Journal of Economic Perspectives 26, no. 4 (2012): 43-56.

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

Huffaker, Ray G., Rodney L. Clouser and Sherry L. Larkin, “Contract for Analytical Services Related to the Deepwater Horizon Disaster: Estimation of Lost Indirect and Passive Use Economic Value to Floridians,” Food and Resource Economics Department, University of Florida, Final Report, March 29, 2012.

Ioannidis, John, and Chris Doucouliagos. "What's to know about the credibility of empirical economics?" Journal of Economic Surveys 27, no. 5 (2013): 997-1004.

Kling, Catherine L., Daniel J. Phaneuf, and Jinhua Zhao. "From Exxon to BP: Has some number become better than no number?" Journal of Economic Perspectives 26, no. 4 (2012): 3-26.

Maas, Harro, and Andrej Svorenčík. "“Fraught with controversy”: organizing expertise against contingent valuation." History of Political Economy 49, no. 2 (2017): 315-345.

McFadden, Daniel, and Kenneth Train, eds. Contingent Valuation of Environmental Goods: A Comprehensive Critique. Edward Elgar Publishing, 2017.

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

Navrud, Ståle. Review of McFadden, Daniel, and Kenneth Train, eds. Contingent Valuation of Environmental Goods: A Comprehensive Critique. Edward Elgar Publishing. Economics of Energy & Environmental Policy, 7(1): 153-158, 2018.

Park, Timothy, John B. Loomis, and Michael Creel. "Confidence intervals for evaluating benefits estimates from dichotomous choice contingent valuation studies." Land Economics 67, no. 1 (1991): 64-73.

Portney, Paul R. "The contingent valuation debate: why economists should care." The Journal of Economic Perspectives 8, no. 4 (1994): 3-17.

Randall, Alan. "What practicing agricultural economists really need to know about methodology." American Journal of Agricultural Economics 75, no. Special Issue (1993): 48-59.

Randall, Alan. "Beyond the crucial experiment: mapping the performance characteristics of contingent valuation." Resource and Energy Economics 20, no. 2 (1998): 197-206.

Train, Kenneth. "Comment on "A revealed preference approach to valuing non-market recreational fishing losses from the Deepwater Horizon Oil Spill" and its "Corrigendum" by Alvarez et al." Journal of Environmental Management (2015).

Whitehead, John, "The re-release of the Exxon-funded RTI contingent valuation critique," The Environmental Economics Blog, click here, November 3, 2010.

Whitehead, John, "What I did on election night," The Environmental Economics Blog, click here, November 6, 2012.

Whitehead, John C. "Plausible responsiveness to scope in contingent valuation." Ecological Economics 128 (2016): 17-22.

Whitehead, John C. "Who knows what willingness to pay lurks in the hearts of men? A rejoinder to Egan, Corrigan, and Dwyer." Econ Journal Watch 14, no. 3 (2017): 346.

Whitehead, John C., "A comment on Desvousges, Mathews and Train (Land Economics 2015): ‘An adding up test on contingent valuations of river and lake quality’." Ecological Economics 177 (2020).

Whitehead, John C., Review of: McFadden, Daniel, and Kenneth Train, eds. Contingent Valuation of Environmental Goods: A Comprehensive Critique. Australian Journal of Resource and Environmental Economics 62(4): 710-713, 2018.

Whitehead, John C., Tim Haab, Sherry L. Larkin, John B. Loomis, Sergio Alvarez, and Andrew Ropicki. "Estimating lost recreational use values of visitors to northwest florida due to the deepwater horizon oil spill using cancelled trip data." Marine Resource Economics 33, no. 2 (2018): 119-132.

Whitehead, John C., John B. Loomis, Andrew Ropicki, Sherry L. Larkin, Tim Haab and Sergio Alvarez, “Estimating the benefits to Florida households from avoiding another Gulf oil spill using the Contingent Valuation Method: Internal validity tests with probability-based and opt-in samples, Applied Economic Policy and Perspectives, forthcoming.

Table 1. Comparison of Whole Scenario Logit Models and Willingness to Pay Estimates

	Desvousges, Mathews and Train (2015)		Chapman et al. (2009)	
	Coefficient	SE	Coefficient	SE
Constant	0.594	0.235	1.071	0.099
Cost	-0.0024	0.0012	-0.0049	0.0005
χ^2 (p-value)	4.29 (p = 0.038)		100.04 (p < 0.01)	
Sample Size	172		1093	
Willingness to pay	434.38		278.63	
	95% Confidence Intervals			
Delta Method lower	102.29		238.60	
Delta Method upper	766.46		318.67	
Krinsky-Robb lower	233.70		244.77	
Krinsky-Robb upper	2062.14		328.21	

Figure 1. Dichotomous Choice Contingent Valuation Bid Curve

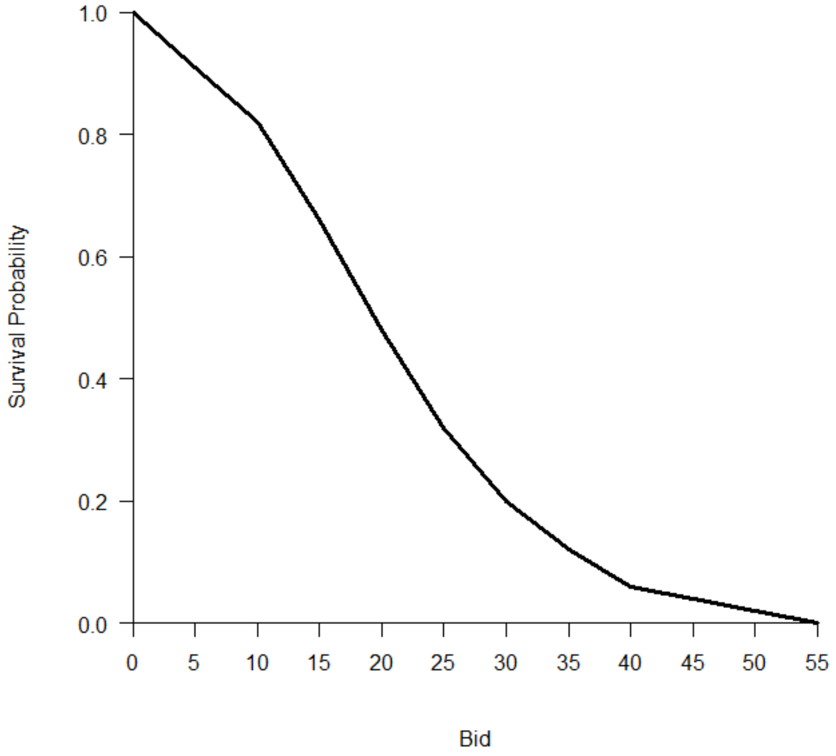
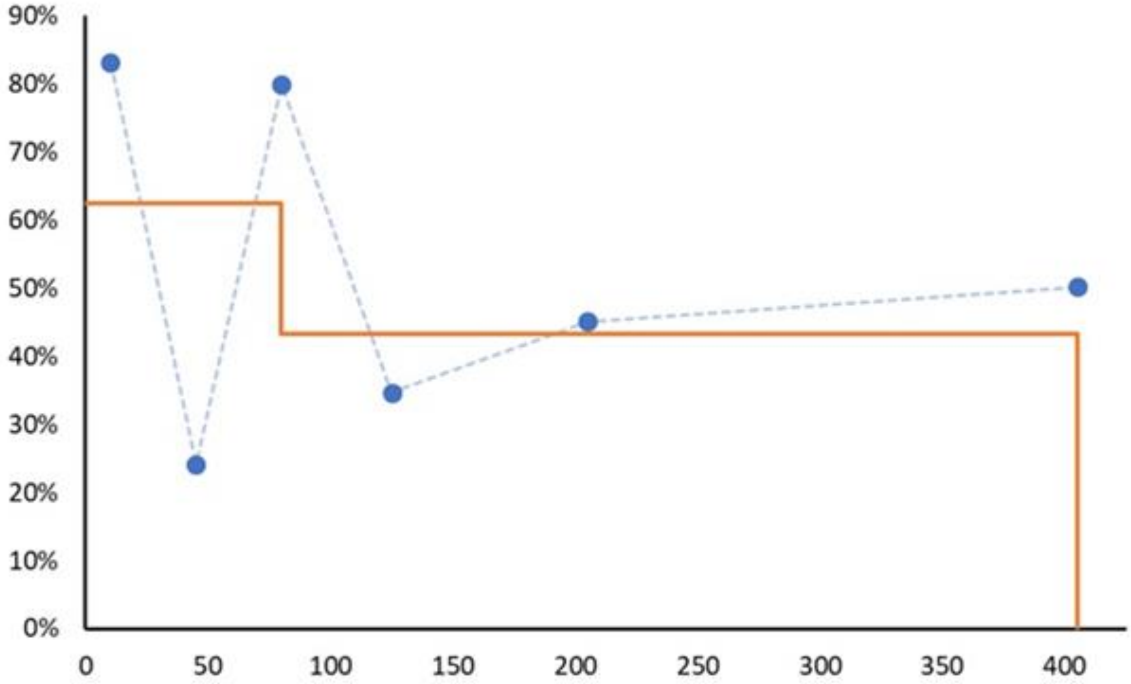
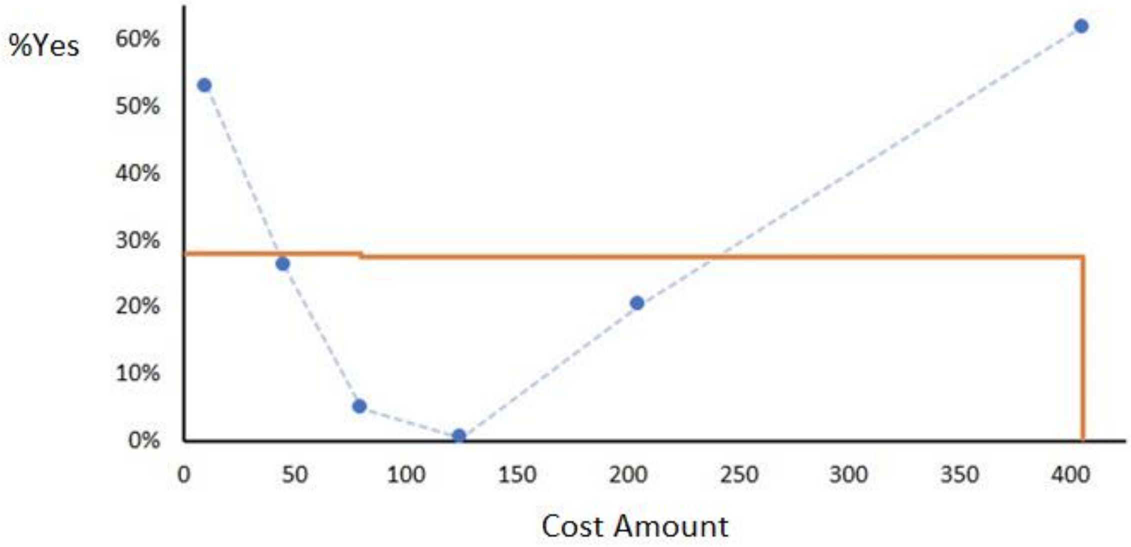


Figure 2. Whole Scenario: Weighted Referendum Votes and Turnbull Curve



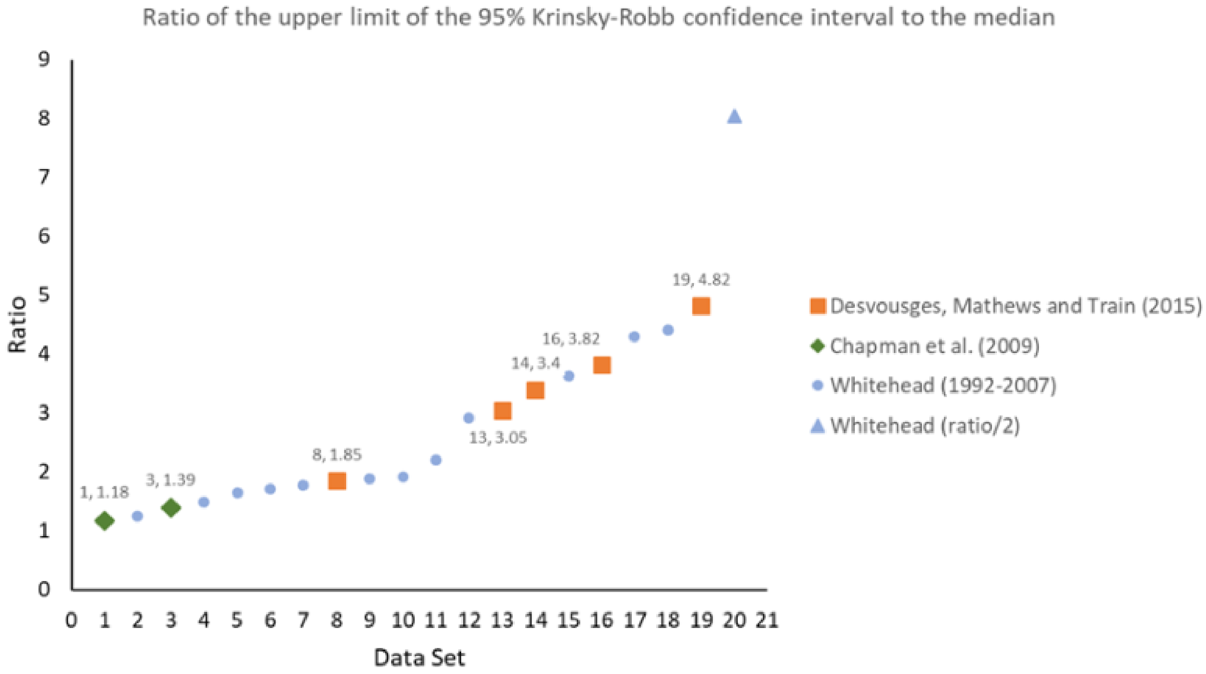
Note: The blue dots are the proportion of weighted referendum votes in favor of the policy at each dollar amount. The orange line is the Turnbull summary of the smoothed data.

Figure 3. Second Scenario: Weighted Referendum Votes and Turnbull Curve



Note: The blue dots are the proportion of weighted referendum votes in favor to the policy at each dollar amount. The orange line is the Turnbull summary of the smoothed data.

Figure 4. Ratio of the upper limit of the 95% Krinsky-Robb confidence interval to the median in 20 data sets


Ratio of the upper limit of the 95% Krinsky-Robb confidence interval to the median
Ratio
Data Set
Desvousges, Mathews and Train (2015)
Chapman et al. (2009)
Whitehead (1992-2007)
Whitehead (ratio/2)