"A replication of willingness-to-pay estimates in 'An adding up test on contingent valuations of river and lake quality' (Land Economics, 2015)"

Response to Referee #1

Referee 1:

This is a nicely done and very thorough replication of a controversial paper which shows that its results are not robust to commonly used alternative econometric specifications. Most of the issues that need corrected with the paper are in the beginning or end of the paper and are mostly to help a reader less familiar with the original papers and topics at issue.

1. The major and easily fixable problem with the paper is that the first paragraph is extremely terse and is not adequate for a reader not closely familiar with the Desvousges, Mathews and Train (2012) paper and Chapman et al. (2016). A little more detail on what Desvouges et al. did would fix this. The argument in Chapman et al. (2016) that the comparison being made is not correct could be put in the current footnote 1 which could be expanded to provide some intuition.

Response: Footnote 1 has been replaced by Appendix A which fully describes the adding test.

The one paragraph introduction has been expanded to five paragraphs that review the literature:

1. Describe the Desvousges, Mathews and Train (2012) study.
2. Describe the adding up test proposed by Diamond (1996) and Desvousges, Mathews and Train's (2012) interpretation of Chapman et al.'s (2009) scope test as an adding up test.
3. Describe Whitehead's (2016) critique of Desvousges, Mathews and Train (2012), the Chapman et al. (2016) comment on Desvouges, Mathews and Train (2012) and Desvousges, Mathews and Train's (2016) reply
4. Describe the Desvousges, Mathews and Train (2015) study.

Appendix A. The adding-up test

Consider two goods, 1 and 2. According to standard scope test theory, willingness to pay for goods 1 and 2, WTP_{1+2}, is expected to be greater than or equal to willingness to pay for good 1, WTP_1. The theoretical construct for the willingness to pay for good 2 used by DMT from Whitehead, Haab and Huang (1998) is a residual difference, $WTP_{1+2} - WTP_1$. Willingness to pay for good 2 in an adding up test is $WTP_2[2|1, Y - A]$ where $2|1$ indicates the valuation of good 2 given the provision of good 1 to account for the substitution effect, Y is income and A is the amount of money taken from the respondent to pay for provision of good 1.

DMT do not explicitly describe the counterfactual situation required by the adding up test. An explicit description of the conditions under which a valuation is made is necessary to account for income and substitution effects. For n goods, the adding up test requires $n + 1$ scenarios

1. Elicit the willingness to pay for good 1
2. Elicit the willingness to pay for good 2 [after describing that good 1 has been provided at a cost of A to the respondent]
3. Elicit the willingness to pay for goods 1 and 2

Following the adding up test theory, in order to accurately elicit $WTP_2[2|1, Y - A]$ one would need to describe the provision of good 1, extraction of WTP_1 and how its provision has reduced the income of the survey respondent before elicitation of WTP_2.

Instead of this additional survey text, DMT elicit WTP_2 just as you would elicit WTP_1 without consideration of substitution and income effects.[3] Inclusion of these two counterfactual conditions in a CVM survey would likely impose additional cognitive burden on the survey respondent.

Referee 1:

2. p. 2, end of second paragraph, not sure that the first part of the last sentence on "fat tails" is the correct description. The main issue is that because the high amounts necessary to drive the percent WTP to zero are not plausible, a good contingent valuation (CV) does not ask them. The problem can be fixed by adding "asked" in front of "does not cause" in the previous sentence and

changing the last sentence to “As such, the WTP estimate will always be sensitive to the assumptions of the estimation approach used.”

Response: The edit has been made:

> This replication is appropriate for several reasons. Dichotomous choice contingent valuation questions propose a cost to respondents who then indicate whether or not they are willing to pay the cost. One theoretical validity test is for whether the percentage of respondents who are willing to pay the cost declines as the cost increases. DMT’s data suffers from non-monotonicity (i.e., the percentage does not always decrease as the bid increases) and flat portions of the bid curve. Another reason for the replication is that the cost range does not cover the entire WTP distribution. In other words, the highest cost amount asked does not cause the percentage of yes responses to fall to zero (Parsons and Myers 2016). As such, the WTP estimate will always be sensitive to the assumptions of the estimation approach used. ~~This “fat tails” problem is pervasive in CVM data (Parsons and Myers 2016) and causes WTP to be sensitive to the estimation approach.~~

Referee 1:

3. p. 2, third paragraph, the ABERS estimator is a “special case” of the “more general” Turnbull (see Turnbull, 1976). Drop the sentence beginning “Both nonparameteric …”. There is only one nonparametric approach (ML subject to a weak monotonicity assumption with variants of the Turnbull depending on the pattern of censoring and truncation with ABERS being the original paper focused on only one of these patterns) and, as is shown in Table 1 of this paper, rarely does the Turnbull approach “obscure” data quality issues. Throughout the rest of the paragraph/paper resolve the ABERS/Turnbull confusion. (Note there is a problem in formula in the Haab and McConnell book in that does not correctly perform the pooled-adjacent low-bound for some patterns of monotonicity violations.)

Response: I have edited this paragraph and refer to the Turnbull throughout the remainder of the paper.

> Following Chapman et al. (2009), DMT choose the ABERS nonparametric estimator for willingness to pay (Ayer et al. 1955). Chapman et al. (2009) describe the ABERS estimator as producing a lower bound WTP estimate. The ABERS estimator ~~is equivalent to the~~a special case of the more familiar Turnbull nonparametric lower bound WTP estimator (Haab and McConnell 1997, Carson and Hanemann 2005, Boyle 2017). ~~Both nonparametric WTP estimation approaches obscure data quality problems.~~ When data is non-monotonic, ~~both ABERS and~~the Turnbull approach~~es~~ smooth~~s~~ non-monotonic bid curves by pooling the percentages of those willing to pay across cost amounts and ignore validity problems associated with flat portions of the bid curve. ~~Both the ABERS and~~The Turnbull estimates truncate the WTP distribution at the highest bid, ignoring the tail of the WTP distribution.

Referee 1:

4. p. 2, change last sentence to "… studies on conducting sensitivity analysis for WTP estimation approaches."

Response: The change has been made:

(2015). In the conclusions I offer recommendations for~~on~~ future CVM studies ~~on~~ on conducting sensitivity analysis for WTP estimation approaches.~~one of these is to conduct sensitivity analysis over WTP estimation approaches when CVM data is "difficult."~~

Referee 1:

5. p. 3, top of page, need a couple more sentences for most readers that explain the nature of the data in Table 1.

Response: Two explanatory sentences have been added:

The data from DMT is presented in Table 1. The cost amounts presented to respondents for each scenario is presented in the first column. The number of "yes" responses, the subsample size and the percentage of "yes" responses is presented for the Whole, First, Second, Third and Fourth scenarios. Each of the scenarios exhibits non-monotonicity in at least one of the five cost increases. For example, in the whole scenario the percentage yes is 61 at $45 and 69 at $80. The whole, first and fourth scenarios exhibit non-monotonicity in the cost increase from $205 to $405.

Referee 1:

6. p. 3, paragraph beginning "Even when …" change "not different" to "not statistically different". This paragraph overall is confusing because the situation being illustrated is not the case where strong monotonicity is meet. Further, what the reader needs to know here although economic theory suggests (weak) monotonicity should hold sampling variability can cause such a constraint to violated in empirical samples (and as this paper notes but at the end, particularly in samples sizes for individual bid points as small as those in DMT).

Response: The change has been made:

Even when the yes responses are monotonically decreasing in the cost amount, the slope is not statistically different from zero in large portions of the bid curves. For example, the whole and second scenarios are characterized by two flat portions of the bid curve. A stylized example is illustrated in Figure 1 where the percentage of yes responses is constant over the lower range of cost amounts ($10 to $125), is downward sloping from $125 to $205 and flat from $205 to $405.

More explanation about non-monotonicity and the reasons for it has been added in the second paragraph below Figure 1:

> Non-monotonicity can be caused by a lack of theoretical validity of the data, a lack of attention being paid to cost amounts by survey respondents or it could be due to sampling variability when small sample sizes are employed (as in Table 1). With non-monotonic data, nonparametric WTP estimators require pooling of yes responses across cost amounts until weak monotonicity is achieved. Weak monotonicity occurs in the data when the percentage of yes responses is equal across bid amounts. When the probabilities for two pooled costs are higher than the next lowest cost the pooling continues until the bid curve is non-monotonically non-increasing in the cost amount. The pooled dichotomous choice data are presented in Table 2.

Referee 1:

7. p. 4, bottom of the page (and footnote 2). It is not clearly why DMT use a bootstrap approach when the parametric standard errors are well-defined and it is hard to see why there should be any substantial deviation from the assumptions need to estimate the probability of a yes response at each bid point. DMT do not describe how they perform the boostrap calculation of the differences but this is a non-standard bootstrap if it correctly takes into account the stratification caused by random assignment to bid points.

Response: I, too, don't understand why DMT uses the bootstrapped standard errors.

Referee 1:

8. p. 7, first sentence of the conclusion. Would change the first sentence to: "While it is not clear that the adding-up test DMT advocate should hold (Chapman et al., 2016; Whitehead, 2016) in this case, this replication of DMT shows that it cannot be rejected under two of three alternatives and commonly used parametric econometric specifications.

Response: This change has been made.

> While it is not clear that the adding-up test DMT advocate should hold (Chapman et al., 2016; Whitehead, 2016) in this case, this replication of DMT shows that it cannot be rejected under two of three alternatives and commonly used parametric econometric specifications. ~~Desvousges, Mathews and Train's (2015) dichotomous choice CVM data leads to WTP estimates that fail to reject the null adding up hypothesis test with two of three alternative parametric estimates of WTP.~~ In addition, the weighted WTP estimates fail to reject the null hypothesis of equality between WTP for the whole and the sum of parts with all three parametric

Referee 1:

9. p. 8, would drop the sentence: "Many of the problems …", as it speculates on the size of DMT's research budget, which was likely large enough to support much large sample sizes.

Response: This sentence has been deleted:

survey, fail to replicate the Chapman et al. (2009) study. Chapman et al. (2009) use in-person interviews with a large probability sample as recommended by Arrow et al. (1993). ~~Many of the problems in the DMT (2015) data may be due to the lack of a large research budget.~~ Researchers who are tempted to use these inexpensive panels with online surveys and small samples should