Response to Referee #3

Overall, I think the paper needs to start with a discussion of adding up and why your analyses are relevant. The paper is very mechanistic with little intuition or conceptual guidance. It is written as a note to a peer with whom you have an ongoing discussion, but to the outsider they will likely ask why is this relevant and how do the empirical results address the relevant issue.

P. 2 – "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 nonmonotonicity (i.e., the percentage does not always decrease as the bid increases) and flat portions of the bid curve." Flat portions of the response function do not violate theoretical validity nor monotonicity, but you seem to accidentally imply that they do in you statement.

Response: Deleted

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

P. 2 – "Another reason for the replication is that the cost range does not cover the entire WTP distribution. In other words, the highest cost amount does not cause the percentage of yes responses to fall to zero."

This is a two-sided issue, either the full "value" range is not covered, or some people will never say no to any bid. You do not have enough information to say which or if both are occurring in the data.

Also, putting a bid in the tail of the distribution contributes little to estimation.

Response: I have deleted this sentence.

Section 2 – I suggest briefly reviewing what DMT valued and the definition of each of the scenarios, so the reader does not need to go back to the original paper to understand these treatments. A fundamental issue in adding up is whether it is accepted that the parts should in fact sum to the whole.

Response: I have added an appendix that describes the Chapman et al. (2009) and DMT (2015) CVM scenarios.

Table 1 – Identify the monotonicity violations with bolding or shading to make them apparent to the reader given the amount of data in the table.

Response: These are now in bold.

Table 1 – An issue that you do not mention when discussing monotonicity violations and flat portions of the response functions is that the number of observations at each bid point is small; in some cases, not large enough for the law of large numbers to apply. This is an issue I am not aware of being discussed in the literature. Is there a need to think about sample sizes by bid amount, not just by the total sample? This has implications for choice experiments and experimental economics as well where bid amount effects are rarely considered.

Response: Small sample size is now mentioned as a problem on page 5:

Estimation of the ABERS and Turnbull requires a valid cumulative distribution function (Haab and McConnell 2002). A valid cumulative distribution function is non-decreasing in the cost amount. In invalid CDF is non-monotonic. Non-monotonicity can be caused by either a lack of theoretical validity of the data, a lack of attention being paid to cost amounts by survey respondents or 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.

Small sample size is discussed as an issue in the conclusions:

The data quality problems are particularly apparent in the whole and second scenarios which are the versions of the survey developed by Chapman et al. (2009). Considering this, DMT (2015), who use a relatively inexpensive, small non-probability opt-in sample ~~and an online survey~~(Bill Desyousges, personal communication, February 19, 2015) and an online survey that lacks evidence of consequentiality (Carson and Groves 2007, Carson, Groves and List 2014), fail to replicate the Chapman et al. (2009) study. Chapman et al. (2009) use in-person interviews with a large probability sample and in-person interviews, as recommended by Arrow et al. (1993), and provides evidence of consequentiality. ~~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 do so with caution (Sandorf et al. 2016).

Future studies should attempt to address these problems with larger subsamples of data. This can be achieved in three ways. The most costly, of course, is simply by increasing the overall sample size. Holding cost constant, researchers could reduce the number of cost amounts used in the experimental design or reduce the number of experimental scenarios presented. For example, DMT could have implemented their adding up test with three (instead of five) separate scenarios as they describe in Desyousges, Mathews and Train (2012) (see also Whitehead 2017).

Figure 1 – Was the DMT replication consequential? If not, then there is no reason for subjects to take the bids seriously and this could lead to the response pattern you show. The Chapman et al. study was consequential.

Carson, R.T. and Groves, T., 2007. Incentive and informational properties of preference questions. Environmental and resource economics, 37(1), pp.181-210.

Carson, R.T., Groves, T. and List, J.A., 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice. Journal of the Association of Environmental and Resource Economists, 1(1/2), pp.171-207.

Response: Consequentiality is now mentioned in the conclusions (see the first paragraph in the "small sample" response above).

P. 5 – "I combine the data from the subsamples and estimate linear and log linear parametric dichotomous choice models as recommended by Boyle (2017)" I do not see this recommendation in Boyle, but see no problem with this approach. However, when pooling data, you do this because correlation in responses across treatments yields estimation efficiency. You may want to think about the structure of the corelated error terms.

Response: "recommended" has been replaced with "described." I'm a bit confused by the second part of this comment. When pooling data across treatments with only one response per respondent, I don't see how the responses and error terms will be correlated. Except for separately estimated constants, the pooled sample logit models in Table 4 are fairly standard in the CVM literature.

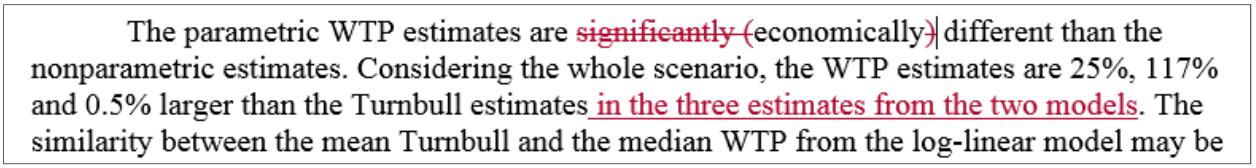
Table 5 – Identify that the t-stats are for whether the estimates are different from zero.

Response: I have added a table note that "t-statistics are for the null hypothesis that WTP is equal to zero."

P. 6 – "The parametric WTP estimates are significantly (economically) different than the nonparametric estimates."

I do not think you want to say "significantly" because you did not do statistical tests, and such test would be questionable because these are not data from independent samples. They are "economically relevant".

Response: I have deleted "significantly":


The parametric WTP estimates are ~~significantly (~~economically~~)~~ different than the nonparametric estimates. Considering the whole scenario, the WTP estimates are 25%, 117% and 0.5% larger than the Turnbull estimates in the three estimates from the two models. The similarity between the mean Turnbull and the median WTP from the log-linear model may be

P. 6 – "The null hypothesis of equality between WTP for the whole scenario and WTP for the sum of the parts cannot be rejected in two of the three adding up tests." Specify these hypotheses, identify the tests conducted and report the test statistics.

Response: I have added the null hypothesis and the test statistics on page 8:

The null hypothesis of equality between WTP for the whole scenario and WTP for the sum of the parts cannot be rejected in two of the three adding up tests. The null hypothesis is , H0: $\sum WTP\ parts = WTP\ whole$ The linear logit that allows for negative mean WTP estimates yields a difference of $168 that is not statistically different from zero ~~as the 95% confidence intervals overlap~~(t=1.12). These WTP estimates pass the adding up test. In the linear logit with the mean WTP constrained to be positive the difference between the whole and the sum of the parts is $680 that is statistically different from zero (t=2.85). ~~The upper limit on the 95% confidence interval for the whole scenario is 766. The lower limit on the 95% confidence interval for the WTP for the sum of the parts is 785.~~ These WTP estimates fail to pass the adding up test. The log linear logit produces a difference of $187 in median WTP that is not statistically different from zero (t=1.05). The median WTP estimates pass the adding up test.

While WTP is the policy-relevant statistic, why not test equivalence of the underlying responses functions (constant + slope)?

Response: I have added an appendix with this analysis for the linear logit (results are similar for the log-linear logit (see results below). The constants are not statistically different for the whole, first and fourth scenarios. The slopes are not statistically different across any of the scenarios. WTP estimates from this model will be equal for the whole, first increment and fourth increment. Naturally, these estimates will not pass the adding up test. Note that this model does not support the results of DMT (2015). These results should be interpreted as data lacking in divergent validity. For example respondents are willing to pay the same amount for cleanup of the river and lake (whole scenario) and cleanup of just the river (first increment). In this comparison the data do not pass the scope test unless the value of the lake cleanup is equal to zero. However, this is contradicted by an identical WTP for lake cleanup in the fourth increment. One conclusion of this analysis is that the data can not support the level of analysis attempted by DMT (2015), especially relative to the more robust results of Chapman et al. (2009) with the same survey.

```
--------------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                VOTE
Log likelihood function      -646.75217
Restricted log likelihood    -679.27607
Chi squared [  9](P= .000)     65.04782
Significance level               .00000
McFadden Pseudo R-squared       .0478802
Estimation based on N =    980, K =   4
Inf.Cr.AIC  =   1301.5 AIC/N =    1.328
--------+-----------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
    VOTE|  Coefficient       Error       z    |z|>Z*         Interval
--------+-----------------------------------------------------------------------
   WHOLE|      .66112***      .10990     6.02   .0000       .44572     .87653
   FIRST|      .66112***      .10990     6.02   .0000       .44572     .87653
  SECOND|     -.28502         .18237    -1.56   .1181      -.64247     .07243
   THIRD|      .19580         .17219     1.14   .2555      -.14170     .53329
  FOURTH|      .66112***      .10990     6.02   .0000       .44572     .87653
 AMOUNTW|     -.00306***      .00052    -5.88   .0000      -.00408    -.00204
 AMOUNT1|     -.00306***      .00052    -5.88   .0000      -.00408    -.00204
 AMOUNT2|     -.00306***      .00052    -5.88   .0000      -.00408    -.00204
 AMOUNT3|     -.00306***      .00052    -5.88   .0000      -.00408    -.00204
 AMOUNT4|     -.00306***      .00052    -5.88   .0000      -.00408    -.00204
--------+-----------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jan 31, 2018 at 04:52:42 PM
--------------------------------------------------------------------------------
```

An important issue you do not address is the fat tails on the distributions. These tails add variance to the estimation that may lead to failure to reject the null of no difference with the parametric estimator. Thus, one could argue that you used the wrong specification and your result is an artifact of your choice of a functional form that does not apply to the data.

Response: One of the three tests, the median WTP from the log-logit, is not sensitive to the fat tail. This test does not reject the null hypothesis of adding up. Note that one of the criticisms of DMT (2012) is that their poor quality data suffers from fat tails so that truncating WTP would lead to more conservative WTP estimates but not provide any more insights into the adding-up test.

P. 6 – "When I apply the same post-stratification weights, scaled to equal the sample size of n=980, to the models in Table 4 and estimate WTP as in Table 5, none of the three sets of parametric WTP estimates supports rejection of the null hypothesis of equality between WTP for the whole and the sum of the parts."

This is true, but it only addresses the whole sample, not samples for individual bid amounts and their inherent response patterns. I am not sure your robustness analysis address fundamental estimation issues and contributes much to the paper.

Response: I have moved the "further tests" section to an appendix.

1st para. of Conclusions – This reiteration of results is not needed in such a short paper.

Response: I have deleted much of the first paragraph:

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 estimates. And using a subsample of data discarding respondents who do not answer the age and income questions, the WTP estimates fail to reject the null hypothesis with all three parametric estimates. DMT's results are not robust to alternative, but standard, parametric approaches to estimating WTP.~~ The failure to replicate DMT's results with the parametric models is due to a host of data quality problems: non-monotonicity, fat tails and flat portions over wide ranges of the bid function. Each of these problems leads to high variability in mean WTP across estimation approach and larger standard errors than those associated with nonparametric estimators.

P. 7 – I concur with the first two paragraphs and think your points can be stated more clearly. I find the last paragraph problematic for the issues raised above and this paragraph takes away from the impact of the content from the two preceding paragraphs.

Response: I have deleted the last paragraph of the conclusions.