Response to Reviewer #2

This is a brief, but thorough and insightful paper. As a profession we should be appreciative of replication efforts like this, especially for methods as important and controversial as contingent valuation. Most of my comments are minor, and are discussed below. But I do have two more substantive suggestions/questions.

First, it is unclear how the statistical comparisons are performed when testing whether the sum of the WTP for each incremental part of the improvement equals the WTP for the total improvement as a whole. In some cases, failure to reject the null hypothesis that the difference of the WTP estimates is zero is discussed, and in other cases it sounds like the author is simply looking at whether the confidence intervals overlap. I suggest providing further details on the statistical test used, and include the test statistics and p-values. This is particularly important for three reasons: (i) the point estimates from the parametric models demonstrate economically significant differences, (ii) the confidence intervals are fairly wide, and (iii) the conclusions based on the upper and lower end of the confidence intervals are fairly close. Poe et al. (2005) discuss alternative methods to statistically compare differences between WTP estimates, including bootstrapping (as done in the original 2015 study by Desvousges, Matthews, and Train (DMT)) and the methods of convolutions, both of which involve simulating a distribution of the differences between the WTP estimates. Poe et al. find that drawing statistical inference based solely on whether the confidence intervals overlap is inappropriate. All this said, since the estimates for each incremental part are estimated within the same logit model, perhaps a more conventional test (e.g., Wald test) was used?

Response: I have provided further details on the statistical test used (page 7) and include the test statistics and p-values (page 8).

Page 7:

> The parametric willingness to pay estimates are presented in Table 5. Mean (and median) WTP from the linear logit, which allows negative WTP, is the negative ratio of the constant and the slope: $WTP = -a/b$ (Hanemann 1984). Estimating WTP only over the positive portion of the distribution from the linear logit uses the formula: $WTP = \left(\frac{-1}{b}\right) \ln(1 + \exp(a))$ (Hanemann 1989). Median WTP from the log linear logit is the exponential of the negative ratio of the constant and slope: $WTP = \exp\left(-\frac{a}{b}\right)$. Mean WTP from the log linear model is undefined when $-\frac{1}{b} > 1$ (Haab and McConnell 2002) as in these models. Standard errors for individual WTP estimates and the sum of the WTP parts are estimated with Wald test and the Delta Method (Cameron 1991, Greene 2017).

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 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.

This brings me to my second question. Why were the parameters in Table 4 estimated using the same regression when each scenario could have been estimated as a separate logit model? The constant term and slope coefficient are identified based on mutually exclusive subsets of the data, correct? Is there any advantage or disadvantage to pooling the data versus estimating separate models?

Response: Estimation within the same logit model makes it easier to conduct the adding-up test. In LIMDEP it only requires a single line of code to produce the mean/median estimate and standard error for each WTP component and the sum of the WTP parts. Estimating the individual models with separate samples produces similar estimates.

More minor comments:

1. A bit more context could be provided in the first paragraph or two of the paper. In particular, a diagram of the four incremental improvements and the whole would be very useful in getting the reader up to speed. Perhaps you can even just borrow Figure 1 from the 2015 DMT paper?

Response: More context (intuition and background) has been added to the introduction and Appendix A in response to referee #1.

2. Although this may be well-understood by most readers, I think a brief sentence in the third paragraph about why the ABERS and Turnbull estimators require smoothing would be appropriate. As I understand it, this is simply because a valid CDF must be monotonically nondecreasing.

This explanation could also fit in the paragraph directly preceding table 2.

Response: I have added an explanation 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.

3. Table 1 would be more legible if there were vertical lines, additional space, and/or alternating background color shades denoting that each set of three columns correspond to a different increment.

Response: I have added outside borders to each of the tables to separate different data, models and estimates.

4. In the conclusion the author refers to DMT's "relatively inexpensive, small non-probability sample". Given the back and forth discussions posted on 9/29/2017 regarding the earlier working paper, it seems that there is some additional information available regarding the internet panel and selection process for respondents. An additional few sentences on these details would be enlightening, particularly if it's believed that the respondents in this particular panel are not fully considering the tradeoffs in the valuation questions. The distinction between methodological shortcomings versus implementation and data quality is crucial.

Response: DMT use Survey Sampling International's opt-in panel. I have used this panel myself several times and, while it provides sometimes valid data at low cost, respondents take less care in the valuation task than researchers hope for. In other, currently unpublished, research I have found that the CVM data passes the scope test with a probability-based sample but not with an opt-in sample, all else equal. Authors who use opt-in panel data should refrain from conclusions without disclaimer about their relatively low quality sample.

I have (a) added a reference to support the sample statement "(Bill Desvousges, personal communication, February 19, 2015) and (b) added a reference to Sandorf et al. (Ecol. Economics 2016), both on page 10.

Works Cited

Poe, Gregory, Kelly L. Giraud, and John Loomis. 2005. Computational Methods for Measuring the Difference of Empirical Distributions. American Journal of Agricultural Economics 87(2): 353-365.