Reviewer's Report:

Chesapeake Bay Watershed Stated Preference Study Draft Report

US EPA

(October 27 Version)

Gregory L. Poe, Ph.D.

Professor, Cornell University

Dyson School of Applied Economics and Management

407 Warren Hall

Ithaca, NY

glp2@cornell.edu

November 25, 2014

Reviewer's Report:
Chesapeake Bay Watershed Stated Preference Study Draft Report
US EPA
(October 27 Version)

The following review is divided into two sections: I) Response to Charge Questions from the US EPA; and II) More Detailed Comments in Chronological Order. The latter section consists of comments and queries beyond the charge questions, related to specific pages, sections, lines and tables in the Draft Report. While many of these specific comments are minor, others address substantive issues that should be regarded as equally important as the responses to the charge questions.

Section I: Response to Charge Questions:

1. *Overall:*
   a. *Does the overall report reflect scientifically appropriate use of stated preference methods and analysis for estimating use and non---use benefits of the TMDLs for the Chesapeake Bay and lakes in the watershed? Is any additional information or analysis needed to characterize the benefits of the TMDLs for the Chesapeake Bay and lakes in the Watershed?*

      The report summarizes a well-done research effort. In my opinion the overall research project was nicely conceived, the survey development followed contemporary standards, proper multi-contact survey implementation procedures were implemented, and attention was paid to potential non-response biases. The baseline analysis is straightforward, utilizing accepted, widely used analytical techniques. With some caveats, the output from this stated preference research should be informative to policy evaluations of the Chesapeake Bay TMDL.

      The core models of this Report, encompassed in the first 43 pages of the text, seem to hold together well and tell a reasonable story. The main effects design and the various interaction terms explored by the authors are informative and make sense. Most of my inputs with respect to these sections are for clarification, and not major.

      Nevertheless, there is one line of investigation that I feel may enhance the understanding of the data: that is while authors take care to pool the Bay States, the Watershed States and Other East Coast States into a single data set, additional insight might be gained by re-estimating a subset of the models (say Models 3 and 6) on each of the three strata. First, the area covered by this survey is large, and there may be differences across strata, states etc. that are correlated with attitudes toward the Chesapeake Bay Watershed and maybe to willingness to pay for the TMDL program. If so, this might change the willingness-to-pay aggregation strategy. This geographical disaggregation is further motivated by my general preference for an approach that builds up data from different samples, systematically testing whether

the data sets belong together, and understanding response patterns and variable interaction in each of my data subsets. Finally, there seems to be an odd result in the Bay-States-Only estimation reported in Table 32 (p. 57) that none of the coefficients of the Chesapeake Bay attributes are significant in the constant Baseline with SQC. Comparing this result with many significant attribute coefficients in Model 1 in Table 11, suggests that Non-Bay-State strata are driving responsiveness to changes in attributes. If I am interpreting these two tables correctly, this result does not make sense to me.

Moreover, I am concerned about variation across strata in the proportion of respondents who agreed that "water quality improvements to lakes outside the Chesapeake Bay" affect their vote. It may be that out of watershed non-users, who seem to have a high annual average household willingness to pay (see table 17), may be responding in part to their more local inland waters or estuaries. Mitchell and Carson, in their 1989 book, refer to this as Amenity Specification Bias in which the respondents are including other attributes or commodities in their willingness-to-pay decisions beyond those conceived by the researcher. Other protest responses, awareness of the Chesapeake Bay and nutrient/sedimentation issues, attitudinal responses, and non-response patterns may differ systematically across strata, which again could affect aggregate results.

While I am positive regarding the main core of the Draft Report, the penultimate section (Section 8 – Validity Tests and Sensitivity Analysis) raises some concerns. Some of these concerns may be due to the presentation, which seems to have been hastily assembled: table references generally do not coincide with table numbering after page 44, I cannot interpret table 27 (which seems to have entries identical to those in Table 30, but in the wrong places), and I am not convinced by a number of the interpretations. Moreover some of the results in this section, if I am interpreting them correctly, seem to undermine approaches used in the core models and methods. I address my concerns about this section of the Draft Report in the following subsections.

b. Please comment on the overall clarity of the report.

As indicated above, I feel that the core of the report, encompassing Sections 1-7 and the first 43 pages, is well presented. My detailed comments in Part II of this review offer additional suggestions and comments.

As indicated in Section I.1a. above, several parts of Section 8 in the report need to be revisited. I provide specifics in my Section I.4c. below.

c. *Does the document appropriately summarize the recent, relevant stated preference literature for valuation of water quality attributes and the Chesapeake Bay?*

I think so. However, it should be made clear that the literature review is focusing on U.S.-based research and estuaries only. For example, there is a large body of contemporary work around the European Community Water Framework Directive that is not included in the review. On p. 9, from line 26 on, the report does extend its focus to inland lakes. There is a body of unmined/undermined literature in that area (e.g. Stumborg, Baerenklau and Bishop, 2001; Johnston and Thomassin, 2010; van Houtven et al. 2014). Since inland lakes are not the primary goal of this report, the authors may not desire to expand in the direction as, I suspect the literature is fairly deep and widely distributed.

Given that there are other studies that have looked at large watershed-level or other water-policy interventions that cover a large area, it may be useful to place your results in perspective of these studies. For example, is the average range of $82 to $167 per household per year large or small? This is a little difficult to get perspective on, as some for the other major estuary studies tend to report willingness to pay in different terms. For example, Johnston et al. (2002) report values expressed in terms of value per acre of different types of land purchased in the Peconic Bay watershed. Kraght and Bennett (2011) report values in terms of values for a hectare increase of sea grass area or a kilometer of native riverside vegetation in Australia. Nevertheless, Van Houtven et al. (2014) report mean annual willingness to pay from lake water quality improvement in Virginia of $60, a bit lower than your lower bound. Metcalf et al.'s (2012) valuation study of the European Community Water Framework Directive in England, reports average household willingness-to-pay values ranging from £50.5 to £268.5. Placing your results in the range of other large-scale studies would also be important because some of the Chesapeake Bay Study design features (as discussed elsewhere in this review) or response factors may tend to push up estimated values: these include the use of DCE instead of CV, not correcting for consequentiality, and evidence of upward omitted variable bias of 35% (p. 58, line 9).

d. *Are the analytical methods used in the Report consistent with the current state of literature? Is the application of these methods appropriate for the stated preference survey data?*

In general yes. The main effects random utility model employed in this research is standard in discrete choice experiment research (see Bekker-Grob, Ryan and Gerard, 2012). The various interactions considered between attributes and location, resource use, and demographic characteristics allow exploration of response patterns.

Yet there are a couple of "current state of the literature" issues that should be addressed in a Final Report, via discussion and/or justification of the choice of approach made. These do not fit neatly into any of your specific queries below. Hence

I will address them here.

Consequentiality: Stated preference methods have long been the subject of the influential critique that hypothetical questions get hypothetical answers. Economics offers little, if any, guidance in how to interpret such responses and many economists correspondingly argue that hypothetical responses are not reliable measures of actual choices.

In the last decade of so, the conceptualization of stated preference survey responses has undergone a paradigm shift (at least by practitioners of the method). In essence this shift has been to reconceptualize survey responses as providing a source of revealed preference information: "As long as the preference information collected in surveys is used by governments and private firms to help make decisions, then individuals *should* use the opportunity provided by their survey response to help influence this decision." (Carson and Groves, 2011, p. 302). The theoretical underpinning of this shift is attributed to Richard Carson and colleagues, funded in part by the US EPA (see U.S. EPA Cooperative Agreement R-824698), who argued that stated surveys should be assessed in terms of their consequentiality and incentive compatibility. For a survey question to be deemed consequential, "survey respondents need to believe, at least probabilistically, that their responses to a survey may influence some decision they care about" (Carson and Groves, 2011, pp. 301-302).

If a survey is consequential, then we would expect the respondents to behave in economically predictable patterns. Hence, it is important that the value elicitation question is incentive compatible in the sense that "participants cannot do better than voting truthfully." (Vossler, Doyon, and Rondeau, 2012, p. 148).

Evidence from a growing body of experimental economic research, framed field experiments, and classroom based exercises suggest that if there is a non-zero probability of influencing a public good provision decision, then survey responses closely reflect actual choices in induced value and homegrown value settings (see Poe and Vossler, 2011 for a summary of this research). This extends to settings in which the information gathered in a survey is advisory and not necessarily binding on the decision maker. In a field CV study, Herriges et al. (2010) further show that those who believe that a survey is at least minimally consequential have statistically different willingness to pay distributions than those who believe that the survey is not consequential. Johnston (2006) shows that a consequential referendum CV survey closely predicts actual vote percentages. Vossler, Doyon, and Rondeau (2012) find like results for a survey using a discrete choice experiment format.

Altogether, the "current state of the literature" points to consequentiality as a critical design element, one that was not addressed in the Chesapeake Bay study under consideration. Perhaps most germane to the present study is the Vossler, Doyon and Rondeau (2012) study. These authors used a Likert scale, ranging from 1 ("not at all")

to 6 (“very strong”) to measure the belief that respondents thought the survey would be “taken into account by public authorities.” They found that when consequentiality was not taken into account, willingness to pay exhibits what the authors refer to as a “modest (positive) bias”, approximately 30 percent higher than payment collected from a parallel real contributions study. For weakly consequential survey responses, for which the Likert scale responses were 4.0 – 4.1 or higher, the willingness to pay derived from stated preference responses was statistically equivalent to the actual contributions data.

On the positive side, the Vossler, Doyon and Rondeau (2012) paper provides positive results for DCE and the methods used here in terms of incentive compatibility. Notably the admonitions to treat each DCE question independently (e.g. p. A-10) enhance the incentive-compatibility characteristics of the Chesapeake Bay Study.

I do not know if there is any way the authors could assert that the present survey is likely to be consequential. On the face of it, the Chesapeake Bay survey would seem to be a very strong candidate for being regarded as consequential: the survey cover is emblazoned with the EPA name and logo, OMB control numbers, and references to Clean Water Act authorization. Anything that could honestly bolster the perception that the survey was likely to be regarded as consequential would, I believe, be very important to countering possible challenges to dismiss the result of the survey because it is based on “hypothetical data.” For example providing a copy of the cover letter if, as suggested in a focus group session (p. A-61), a “strong cover letter” was actually employed. The authors should (re)read the Vossler, Doyon and Rondeau (2012) paper and assess their survey design relative to the consequentiality/incentive compatibility presentation in that paper.

I am not sure if the following is covered by “any additional information and analysis needed.” (see Charge 1.a.), but an ideal way to explore this issue would be to send out the same survey to a small sample and include an additional consequentiality question in or near Q10. I add that such a follow-up might be deemed very useful to any future stated preference survey research conducted by the US EPA.

<u>Dichotomous Choice Contingent Valuation (DCCV) versus Discrete Choice Experiments (DCE)</u>: While in recent years DCE approaches have dominated the literature there remains a tension between DCCV and DCE about which would be more appropriate to use for valuing a specific policy. The tradeoffs are well summarized in the introduction to a water quality valuation paper that authored by leading stated preference scholars Richard Carson and Ian Bateman and others (Metcalf et al., 2012):

> “The discrete choice experiment (DCE) framework … is naturally suited to the development of multi-attribute valuation models of the kind required in our study. A number of studies have found, however, that the DCE approach focused on multiple policy changes often elicits higher values for the same

> package of improvements than a contingent valuation (CV) study focused on a single policy change... A number of reasons have been put forth for this finding ranging from strategic behavior to placing less weight on the cost attribute when it is varied simultaneously with other attributes to various types of learning behavior. Most of these suggest that the relative values of noncost attributes derived from a DCE can be considered reliable, but that total values, which depend on cost, may be biased upward." (p. WO3526)

In their subsequent analysis of willingness to pay for water quality improvements, the authors report DCE values that are more than twice the DCCV values.

I am not trying to say that the US EPA should have done a DCCV study. What are needed, however, are a clear discussion of why the DCE approach was chosen and the consequences of that choice. I surmise that it was desired to have a valuation method that allowed you to consider a range of possible TMDL outcomes and independent changes in attribute levels? This is credible, and needs to be presented.

Calculation of Variance in Marginal and Total WTP: As far as I can tell, the authors do not describe how the variance of WTP and marginal WTP was calculated. It seem that an analytical approach or delta approach was adopted? Because of the possibilities of non-linearity in ratios of coefficients (see Hanneman and Kanninen, 2001), many authors instead use bootstrapping methods for estimating and comparing WTP values from DCE surveys (e.g. Johnston et al. 2012; Rolf and Windle, 2012)

2. *Is the survey development process clearly described and consistent with best practice in the economics literature?*

My assessment is that the survey development process is consistent with the best practices in the economics literature. The survey was developed with guidance from leading scholars in applied environmental economics (M. Cropper and A. Krupnick) and policy-oriented survey research (E. Besedin) and reviewed by three highly respected stated-preference researchers (K. Boyle, R. Johnston, and J. Whitehead). The actual development of the survey made use of 10 sequenced focus groups and a number of cognitive interviews. From the information presented in the Appendix, *it appears* the information gathering and design process sought to update and incorporate information from preceding steps.

I use the phrase "it appears" in the previous paragraph because I think that the survey development process could be more clearly described in the text and the appendix. The overall presentation would benefit by briefly detailing what changes were made after each set of focus groups. This would strengthen positive external evaluation of the survey. As an example, I suspect that the statement in the discussion in Focus Group 8 (p.

A-58) that "The independence of freshwater and bay water attributes must be further emphasized in the text" led to the inclusion of the "For example: A pollution reduction program close to the bay would improve the water quality in the Chesapeake Bay itself but would not have much affect on the Watershed Lakes." Having a greater link between the main text and the appendix would be useful.

Further, in looking through the Appendix it appears that there were persistent issues and that a particular survey-design decision was made. However, little information is provided about how these issues were resolved. With respect to this concern the factors in the following key design decisions should be elaborated on:

1. Endpoint versus input approaches.
2. The decision not to use water quality indices.
3. The selection of dollar value ranges.
4. Payment vehicle length of time.

(I do note that some endpoint commentary was provided in the text, but I would like to see that more linked to the focus group results.)

I further think that the experimental design could be expanded a bit. While the main effects design seems to be the workhorse of DCE (e.g. see Bekker-Grob, Ryan, and Gerard, 2012), it would be nice to have a justification for that decision. For example, were possible inland lake and Chesapeake Bay interactions considered? Maybe it was just deemed a defensible standard default (which is okay). Also, periodically the issue of two versus three options bubbles up in the literature. While three seems to have emerged as a standard, Vossler, Doyon and Rondeau (2012) seem to advocate the two-option approach based on clarity in decisions rules and incentive compatibility. As I recall, Kate Carson and others (2014) indicate that using more than two options induces status-quo bias. Regardless, a few lines on that decision would be appreciated.

Despite following Dillman-style best practices, the response rates of 0.271 to 0.341 seem low. I am concerned that this may be a result, in part, of contemporary widespread negative perceptions of the US EPA and (environmental) regulations. But my knowledge about policy-relevant survey response rates may be a bit lagged. Your rates should be placed in the context of other recent policy-driven valuation research, preferably by federal agencies. I stress policy-driven because methodological studies can have much lower rates and still be acceptable.

3. *Is the description of the data clear, including summary statistics and comparisons? Are the conclusions drawn consistent with the data?*

Yes, in sections 1-7. No in section 8 (see charge 4c below).

4. *Data Analysis:*
   Considering the main objective of estimating the benefits of the TMDL:
   a. *Are the analytical specifications, variables, and overall methodology consistent with current state of the relevant literature, and appropriate for the survey data?*

      As noted elsewhere, I believe that the analytical specifications in sections 1-7 are appropriate. I would suggest further investigation and estimation of selected models (e.g. models 3 and 6) for each strata independently as discussed in sub-section 1a. above.

      It would be nice to have an expansion of footnote 4. Given the finding of quadratic relationships in some of the attributes in Table 24 (p. 50), I would suggest further discussion and explicit exploration of non-linear specifications (e.g. natural logs). My sense is that the authors have a strong prior to use a linear approach, perhaps to support the use of alternative baselines? I think that the issue of linearity versus non-linearity is critical to extending these results to the baseline scenario described in Footnote 11 and Table 16. If there are non-linearities, then adjustment in the process of estimating benefits from alternative baselines will have to be undertaken.

   b. *Do the results sections provide an adequate, robust, and clear description of the findings from the various models and analytic choices?*

      Based on what is presented, the robustness worries me. It is disturbing that the Table 11 model 1, which shows significance for attributes using data from all three strata does not carry over to the Bay State model (Table 32, p. 57). If I am understanding the two tables correctly, the authors need to explain what is going on: is all the responsiveness driven by the other Non-Bay-State strata? Why is there non-responsiveness in the Bay State data?

   c. *Are the interpretations and conclusions drawn from the analyses consistent with the empirical findings and the relevant economics literature?*

      There are several interpretations and conclusions that I do feel are not consistent with the empirical findings.

      Specifically:

      p. 44, lines 37-39 indicate that "the coefficient estimates are all of the expected sign and are often statistically significant, thus providing support that the results pass the external scope test". First, if something is not significant, it is not significant. Being positive and not statistically significant does not support a scope test. Indeed, it seems to suggest that there is no scope. Second, I count eight of 15 coefficients to not be significant in Table 21, p. 45. Here the glass seems to be half empty, but the

authors interpret it as half full. What these results seems to show is that, aside from the Bass coefficient, evidence for negative scope sensitivity is not found.

p. 48. Table 22. It is not clear how you tested this and what the chi-squared test does. I would think that some form of a likelihood-ratio test, restricting the coefficients to be equal across equations would be appropriate. Of course, you have to determine whether you are holding all the other coefficients equal across questions. For example, if I compared the likelihood ratios of a restricted model for clarity (coefficients not allowed to vary) versus an unrestricted (coefficients allowed to vary) do I assume that all the other coefficients (e.g. bass, crab, oysters, lakes, and cost) are restricted to be equal or not.

p. 48, line 17-20. I personally don't see the "weak evidence of a decreasing marginal utility of income" in Table 23, p.49 (which is referred to as Table 25 in the text). It looks to me like marginal utility of income slightly rises and then falls across income groups.

The results in Table 24. p. 50 do not support the use of linear models elsewhere in the Report. This is problematic.

Table 27 does not support the omitted variables discussion. This table seems incorrect. My sense is that some of the cell entries are not in the correct place, and that the values themselves appear to be some permutation of values in Table 30. For example:

Variable Table 27 Model 1	Coefficient	Variable Table 30, Constant Baseline with SQC	Coefficient
bass X food	4.127e+00***	Status Quo (std dev)	4.127e+00***
crab X food	2.266e-01*	Clarity	2.266e-01*
Status Quo (mean)	8.755e-03***	Crab	8.755e-03

This error is unfortunate from the perspective of this review, because I cannot assess the validity of the conclusion on p. 58, line 9-10 that "There is evidence of omitted variable bias with willingness to pay falling by 35%". This could be an important result.

p. 58, lines 11-13. Finally the WTP estimates estimated from different baseline versions were not statistically different from each other but "nominally showed the reverse relationship". Again, if it is not significant, one cannot interpret directional effects.

d. *Is the non---response data analysis appropriate and consistent with current economic practice?*

Overall, I like the non-response approach and the use of this information to estimate aggregate WTP. However, I think there are a couple of issues that merit attention.

- What was the response rate to the non-response survey?
- Break out Table 6 by strata for both the Main Survey and the Non-Response Follow Up if there is enough sample.

5. *Are relevant issues identified and addressed appropriately in the sensitivity analysis, given the current state of the literature?*

I think the issues identified are appropriate: preference consistency, scope, constant marginal utility of income, non-linearities, screening criteria, omitted factors, and baseline comparisons. However, as noted above in subsection 4c, I think that the analysis, interpretation, and presentation of this section need to be revisited.

6. *Total WTP*

a. *Is the approach to extrapolating household willingness to pay to populations consistent with precedent and best practices in the literature?*

The calculation of household willingness to pay from a weighted regression model and aggregating to the total population is common form for these types of analyses. However, most studies typically cover a smaller geographical area, at say the state level. For example, the van Houtven et al. (2014) and Banzhaf et al. (2006) studies referenced in the report aggregate up the Virginia and New York state levels, respectively.

As raised at various points in this Report, I am worried the present study extends across several states and a wide geographical range. Residents in the above referenced Virginia and New York studies, are likely to be more homogenous than say, residents of Massachusetts and South Carolina. This becomes a concern if omitted characteristics or other variables vary systematically across states and affect willingness to pay.

At a minimum, I think that the three strata should be separated an explored more fully than is done in this report. Perhaps state effects can be added? The basic guiding principle would be starting from a basic presumption that at least the strata (and more likely the states) are different, when can it be shown that they are not, and when can the different subsets of data be pooled. Yes, this will complicate the estimations of aggregate willingness to pay and it may have no

overall effect on average. Yet, such a result, if it exists needs to be demonstrated I believe.

Aggregate values are a product of average household value and population. Because of the population effect and the fact that non-use values may not degrade over distance for an iconic good, the authors need to be clear why only the 17 states within 100 miles were considered. This is not a political boundary, which is the rationale used in state-level aggregation such as Van Houtven et al. (2014) and Banzhaf et al. (2006). Hence the rationale needs to be provided.

b. Is the approach for estimating total WTP for users and non---users consistent with current scientific practice?

I like the non-use value approach. Moreover, if a resource is truly "iconic", I would expect that there is likely not going to be distance effects. However, the evidence presented in Table 6, suggests that awareness of the Chesapeake Bay, and the threat to Chesapeake Bay from sediments and nutrients, is not 100 percent. In the Non-Response Follow Up, only 85% of respondents had heard of the Chesapeake Bay. Hence, while non-use values may remain constant for a fully informed/aware population, the level of awareness about the resource may vary. This could affect the non-use values. To the extent that awareness varies with distance from the Chesapeake Bay, the aggregate non-use values could vary between Bay States and other states for example.

Again, this points to the desirability of trying to look at the data in a more disaggregated form.

7. *The appendices provide additional information and background. Is there an appropriate balance between information presented in the report and information reserved for the appendices?*

Yes. However, as detailed in Section 2 above some additional information needs to be included in the appendices.

II. More Detailed Comments, in Chronological Order

1. Introduction:

p. 5., line 19. I think that advantage of stated preferences are not limited to being the only method to capture nonuse values. I suggest that you stress here, and in the closing lines in the conclusions section on p. 59, that many of the water quality attributes anticipated by the full implementation of the TMDL have not been seen for decades, and hence could not be captured in revealed preferences such as travel cost model (unless you

combined this with stated preferences) or hedonic property valuation. Maybe bring up some of the discussion that appears in the third paragraph of p. 9.

2. Theory and Background:

   p. 6, lines 24-32. Here is where I think that the consequentiality discussion I raised in Section I is relevant.

   p. 6, line 31-32. Do you have any support for the statement "there is general acceptance"? My sense is that if another major non-use value damage case arises that the academics will once again be split into opposing camps on the validity of state-preferences to measure non-use values.

   p. 6, line 37. Louviere and Carson (2011) does not appear in your references.

   p. 7, line 6. You might want to clarify that you are using a discrete choice experiment (DCE) and explain briefly what that is.

   p. 7, line 23-24. This last sentence, and indeed the entire paragraph, need to be linked better. Your sentence oscillates between CE, CV, and Benefit Transfer.

   p. 7, lines 36-37. This last sentence seems to be an add on. I don't understand it.

   p. 8, lines 37-38. What did Windle and Rolfe find with respect to non-users? Is it relevant to your work? My recollection is that they found no statistically significant degradation in non-use values as distance to the Great Barrier Reef increased? This is likely because of the iconic nature of the Great Barrier Reef. I am not sure if the Chesapeake falls in that same category as you seem to imply on the next page. But I admit, I may be wrong in that assessment.

   SP applications for Estuaries and the Chesapeake Bay: It seems that this section is devoted to U.S. studies, which is fine. But that should be noted, as you are missing a lot of the European work in particular on the European Community Water Framework Directive and some related work in AU.

   p. 10, Line 19; I do not understand "values for non-users can be considered a lower bound on non-use value for these improvements." I regard zero as the best lower bound.

3. Survey Instrument Design and Development

   p. 11, line 11. Any more information as to why all states (17) within 100 miles of the east cost of the U.S. was deemed the relevant population? This is important because of the aggregation issues. The number of households you aggregate over will greatly affect your total value estimates so this needs to be clearly vetted.

p. 11, line 35. “Randall” not “Randal.”

p. 12, line 6. It would be helpful here to indicate that there are three baseline conditions.

p. 16, lines 29-34. Is the pretest referred to here the same as the pilot on p. 10 line 37. If so, choose one term.

4. Sample and Data Collection.

p. 17, lines 1-5. This follows a paragraph on the non-response survey. It is not clear if lines 1-5 refer to the non-response survey or actual survey. I suspect it refers to a second $2 given to the non-respondents based on line 6 p. 22.

p. 19. “self-addressed”? Was this addressed to the US EPA, Abt, or? This is related to the issue of consequentiality raised in Section I of this review.

5. Data Summary.

P. 23, Table 6. It would be useful to break out “Heard of Chesapeake Bay” and “Visited the Bay” between sample strata. This could be important if someone is substituting, say, the Peconic Estuary for Chesapeake. It is also critical because variation in these variables may be correlated with willingness to pay cross strata. Again this turns to concerns about aggregation across strata raised elsewhere in this review.

6. Data Analysis

p. 25. After line 20. This is where a discussion of how variance was calculated would be useful.

Footnote 4. See discussion in Section I.4a. in this review.

p. 26, line 29. “AAPOR3”?

Sample Weights: I am not an expert in weighting, so the following question may be naïve. Do the weights have to be adjusted after screening for the various protests, warm glow, and hypothetical biases? If not, please provide support for this conclusion. If so, then please indicate how the changes were made. Regardless of yes or no, some discussion should be put in regarding this issue.

p. 31, line 19. Is this p value for a null hypothesis test of all five user interaction coefficients? Please make a bit clearer as to what is being tested jointly and the results of these tests by adding the degrees of freedom and the chi square test value.

Table 11 and other tables. Indicate what asterisks mean.

p. 32, line 12-13. Here you are referring to Model 3. Is this users or non-users? Moreover, you suggest that the difference is not significant, but you have a p value of 0.06 < 0.10. Elsewhere (e.g. Table 11 and footnote 13) you seem to use a significance criteria of p=0.10. I must be missing something here?

Table 13 Model 5 and discussion on p. 35, Model 5 vs. Model 4. How come when you add education, and not race, as interaction terms all the main effects for Chesapeake Bay become insignificant?

Table 13 Model 6 and discussion on p. 35. Then, when both education and race are included as interaction terms, many of the main effects become significant again? Why does the report add education and then race rather than the reverse? Or consider a model with just race?

P. 35, line 15. You should present and justify your preferred model at this point and elaborate/defend this choice.

Table 14. Why are some of these numbers slightly different (e.g. Main Survey Heard of Chesapeake Bay) or more than slightly different (e.g. NRFU, Aware of nutrient and sediment pollution) than those reported in Table 6? If these are not typos, then explain in text. Also, the authors should investigate alternative statistical tests commonly used on Likert Scale responses such as Kruskal and Mann-Whitney tests. These tests look at patterns in the entire distribution, which seems more appropriate than the pairwise t-tests employed in this Report.

Table 14 and p 38 lines 6-7. As noted Likert Scale response pattern comparisons should involve more that bin frequency t-tests of means. If these tests are significant, you discussion on lines 6-7 could be strengthened.

p. 37, line 9. While this may be just an easy out, this response pattern is consistent with lower rates of participation in recreational activities as well as lower response rates.

7. Estimating Household and Total WTP.

Throughout – you need to be clear about what baselines and projected baseline your are referring to. I assume the constant baseline?

p. 38, fn. 11. You should use “discrete” choice experiment rather than “conjoint.”

p. 39, line 12. Holmes and Adamowicz (2003) is not in the reference list. I presume it is their chapter in the Champ et al. book.

p. 40, lines 4-6. I am not sure exactly what comparisons are being made here: is it users to non-users? holding watershed condition constant or is it in and out of the watershed? holding use constant? (I suspect the latter based on footnote 13). A couple of lines of clarification would be useful. Also, this is where knowledge of how the variance was calculated would be useful.

p. 40, Table 17. Why are Out of Watershed Nonuser Values so high? This should set off alarms! Are they paying for Tar Pamlico or the Passaic River? Results like this suggest a need for greater disaggregation.

p. 40, equation 7. By using Model 3 as an example calculation you are assuming that the Marginal Utility of Income is constant cross income groups. This isn't consistent with Table 18 and subsequent analyses of the various models. Hence, it would be helpful to clarify on or around p. 41 line 2 that you will allow the Marginal Utility of Income to vary when the associated model calls for it.

P. 43 equation 8 and equation 8. You have two equation 8's. Moreover, the superscripts high and low seem to be reversed. I.e. when you adjust by response rate (assume everyone else is zero), you should have WTP^{Low}?

8. Validity tests and Sensitivity Analysis.

p. 44, Line 19. "Attributes", not attribute.

p. 44, line 31. I do not have access to the Carson chapter your refer to. It seems odd to call something an external scope test when you are looking at coefficients to responses within a sample and not across samples. Since I can't verify myself, please double check to make sure this is what Carson was aiming at.

p. 44-46. Table numbering skips Table 22, but refers to Table 22 in text.

p. 44, lines 38-39. I count eight non-significant non-cost coefficients and seven significant non-cost coefficients in Table 21. I don't see how this supports the external scope test?

p. 47, Table 23. Given the positive coefficient for Clarity, I don't understand why the central value estimate for 3 to 3.5 feet is negative?

p. 48. Table 22. It is not clear how you tested this and what the chi-squared test does. I would think that some form of a likelihood ratio test, restricting the coefficients to be equal across equations would be appropriate. Of course, you have to determine whether you are holding all the other coefficients equal across questions. For example, if I were to compare the likelihood ratios of a restricted model for clarity (coefficients not allowed to vary) versus an unrestricted (coefficients allowed to vary) do I assume that all the other coefficients (e.g. bass, crab, oysters, lakes, and cost) are restricted to be equal or not.

p. 48, line 17-20. I personally don't see the "weak evidence". It looks to me like it rises and then falls across income groups.

p. 48, table 24. See comments in Section I.4a.

Consideration of omitted variables. I do not understand the two columns in Table 27, p. 53. The first five coefficients in Model 1 replicate the original Model 1. Beyond that I do not understand why both columns have the interaction. Are there separate number of observations in each of the models (if so, report these)? Also, perhaps more fundamentally, as discussed in Section I.4c. numbers in these tables appear elsewhere.

Comparison of Constant, Declining and Improving Baseline Versions (and p. A-21 and p. A-37): I don't know how you could work around this, but in your focus group discussion, you indicted that some respondents did not treat the attributes independently. Here it seems that all your changes were correlated between relative to the baseline, undermining the desire for independent consideration of each of the attributes.

p. 55, line 5. "from" not "form." (I probably have several of these errors in this review!)

p. 55, line 17. Again, if it is not significant, I would drop the "at least nominally phrases".

p. 57, Table 32. Why are the constant baseline with SQC Chesapeake Bay coefficient attributes not significant. This does not seem to be consistent with expectations and Model 1, in Table 11.

p. 57, line 5 should be "marginal" utility.

p. 59, line 3 should be "willingness to pay on average between $82…"

References:

Banzhaf, H.S., D. Burtraw, D. Evans, and A. Krupnik, 2006. "Valuation of Natural Resource Improvements in the Adirondacks." *Land Economics* 82:445-464.

Carson, K.S., S.M. Chilton, W.G. Hutchinson and R. Scarpa, 2014. "Is Status Quo Bias Design-Induced? Rethinking the Role of Design Selection in Choice Experiments." Paper presented at W-3133 annual meetings, Orange Beach, Alabama.

Carson, R. T., and T. Groves, 2007. "Incentive and Informational Properties of Preference Questions." *Environmental and Resource Economics*, 37:181-210.

Carson, R. T., T. Groves, and J.A. List, 2014. "Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice." *Journal of the Association of Environmental and Resource Economists*, *1*" 171-207.

De Bekker-Grob, E.W., M. Ryan, and K. Gerard. 2012. "Discrete Choice Experiments in Health Economics: A Review of the Literature." *Health Economics*, 21:145-172.

Hanemann, W. M., and B. Kanninen, 2001. "The Statistical Analysis of Discrete-Response CV Data." In I.J. Bateman and K.G. Willis eds. **Valuing environmental preferences: theory and practice of the contingent valuation method in the US, EU, and developing countries**. Oxford University Press, pp. 302-441.

Herriges, J., C. Kling, C.C. Liu, and J. Tobias, 2010. What are the Consequences of Consequentiality? *Journal of Environmental Economics and Management*, *59*: 67-81.

Johnston, R. J., Grigalunas, T. A., Opaluch, J. J., Mazzotta, M., and J. Diamantedes, 2002. Valuing Estuarine Resource Services Using Economic and Ecological Models: the Peconic Estuary System Study. *Coastal Management*, *30*: 47-65.

Johnston, R.J., E.T. Schultz, K. Segerson, E.Y. Besedin, and M. Ramachandran, 2012. "Enhancing the Content Validity of Stated Preference Valuation: The Structure and Function of Ecological Indicators." *Land Economics* 88: 102-120.

Johnston, R.J. and P.J. Thomassin, 2010. "Willingness to Pay for Water Quality Improvements in the United States and Canada: Considering Possibilities for International Meta-Analysis and Benefit Transfer." *Agricultural & Resource Economics Review* 39: 114.

Kragt, M. E., and J.W. Bennett, 2011. Using Choice Experiments to Value Catchment and Estuary Health in Tasmania with Individual Preference Heterogeneity*. *Australian Journal of Agricultural and Resource Economics*, *55*:159-179.

Metcalfe, P.J., W. Baker, K. Andrews, G. Atkinson, I.J. Bateman, S. Butler, R.T. Carson, J East, T. Gueron, R. Sheldon, and K. Train, 2012. "An Assessment of the Nonmarket Benefits of the Water Framework Directive for Households in England and Wales." *Water Resources Research* 48(3) (accessed on line).

Mitchell, R.C. and R.T. Carson, 1989. **Using Surveys to Value Public Goods: The Contingent Valuation Method.** Washington DC: Resources for the Future.

Poe, G. L., & Vossler, C. A. (2011). Consequentiality and contingent values: an emerging paradigm. In J. Bennett Ed., **The International Handbook on Non-Market Environmental Valuation. Cheltenham***,* UK: Edward Elgar Publishing, pp. 122-141.

Rolfe, J., and J. Windle, 2012. "Distance Decay Functions for Iconic Assets: Assessing National Values to Protect the Health of the Great Barrier Reef in Australia." *Environmental and Resource Economics,* 53:347-365.

Schaafsma, M, R. Brouwer, I. Liekens, and L. De Nocker, 2014. "Temporal Stability of Preferences and Willingness to Pay for Natural Areas in Choice Experiments: A Test-Retest." *Resource and Energy Economics*.

Stumborg, B.E., K.A. Baerenklau, and R.C. Bishop, 2001. "Nonpoint Source Pollution and Present Values: A Contingent Valuation Study of Lake Mendota." *Review of Agricultural Economics* 23:120-132.

Vossler, C. A., M. Doyon, and D. Rondeau, 2012. "Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics*, *4*:145-171.