RESPONSES TO CHARGE QUESTIONS

1) Overall

a) Reflect scientifically appropriate use? Additional info or analysis?

In general this report describes a study that was executed according to commonly understood best practice approaches for stated preference analysis. Examples of the steps taken, which a credible contemporary study needs to do, include:

- A systematic survey development protocol which included peer input, focus groups, cognitive interviews, and a pre-test. I particularly liked how the study authors solicited meaningful advice from SP experts at the front end of the study via the RFF workshop.
- Use of a survey vehicle that provided multiple opportunities for identifying protest responses, and responses that may induce hypothetical bias.
- A convincing protocol for sample triage in order to screen out responses that might induce biased estimates.
- Inclusion of a non-response analysis that can help identify sample selection issues. The ability to document differences in attitudes between respondents and non-respondents is a particular strength of the study.
- Investigation of difference sources of heterogeneity in responses.
- Execution of validity and sensitivity analyses.

Attention to these steps implies that the study meets the necessary conditions for its credible use in policy analysis. This said, I do think there are additional things that could be done to improve the usefulness of the report for gauging the validity of the estimates. I will provide specifics on this below, but in general the report is fairly terse in its description of what was a major survey design and execution process. I think additional description on how the focus groups and cognitive interviews informed the process by which the survey evolved would help the reader better judge the process – e.g. something that translates the detailed accounting of the focus groups in the appendix into a summary of what was learned and how it was applied. Likewise, there are additional data summaries and models that could better illustrate the important features of the data.

b) Clarity of report?

By and large the report is nicely written. The sections flow together well and the prose is clear and tight. The style and tone are appropriate for a document of this type – i.e. technical as needed, with simple terms and sentence structure used in the narrative.

The one major exception to clarity lies in how the model and estimation approach are described. There is not much discussion of the type of model estimates (i.e. conditional logit, mixed logit, etc.), though one is left with the impression that estimation will follow the simple and transparent conditional logit paradigm. Upon reading the first and subsequent tables, however, one sees reference to the 'mean' and 'standard deviation' of the status quo ASC parameter, suggesting there is some mixed logit modeling going on. If so, more needs to be provided on estimation approach (simulation with Halton draws?), robustness to difference starting value choices (this can matter when there are a lot of parameters, as with the heterogeneity models), and whether or not the estimates are cross-section or panel (is the random draw the same across choice tasks for a given person?). More generally, the report should include additional

details on how the standard errors were calculated. Are these sandwich estimates? Is any attention paid to clustering, based on multiple choices per person? Knowing this is critical for gauging the precision of the parameter estimates and hence the WTP predictions.

c) Summarize recent/relevant literature for water quality attributes and Chesapeake Bay?
This is another example of where the report should be expanded. The existing literature is mentioned, but little analysis or summary is given. It would be useful for gauging the convergent validity of the estimates from *this* study to know a bit more about what *other* studies found regarding preference for similar water quality attributes. The SP literature on water quality values is comparatively large, and so it would be useful to know:

- What attributes are valued in studies that bear similarity in design and study quality?
- How do the household level WTP findings reported here fit into the range of values found in other water quality studies? Are differences sensible given the different geographic scales of the different studies?
- What are the features of the existing studies that prevent an approximate assessment of the value of Bay improvements based on their findings?

To me these questions are not so much about checking the literature review box, as they are about seeing how the decisions made in the design of this study were taken based on how the specific needs of the Chesapeake differed from other areas. In addition, level estimates of household and aggregate WTP can be hard to interpret without some additional context; understanding how other studies have valued similar/different water quality attributes at similar/scales can provide this. Finally, while it is a far from perfect validity concept, seeing convergence with other well-executed studies (or rational divergence) can bolster the case for the accuracy of the new findings.

d) Are the analytical methods consistent with current state of literature?
This seems to be the case. The approach taken of beginning with a base model, and then progressively adding additional richness, is consistent with common practice. Likewise the systematic approach to validity and sensitivity analysis follows a familiar pattern.

The one exception I will point out was mentioned above. There are choices one can make on how to calculate standard errors, and details on this were lacking. The report should be clear and explicit on the method, and why the method is preferable given characteristics of the sample. This is particularly important, given that many of the estimates are insignificant or only marginally significant.

2) Survey development process clearly described and consistent with best practice in economics literature?
I am quite confident that the survey was developed in a way consistent with best practice. Indeed, the efforts to solicit expert input, and the impressive volume of focus groups and cognitive interviews, suggest that on these dimensions the development is of higher than typical quality. I commend the investigators for such a careful, thorough development process.

In terms of description, I think the report should include more details on how this impressive process was used to evolve the survey instrument. The appendix contains a detailed description of the focus group and

cognitive interview activity, but it is hard there to see the forest above the trees. I would like to see a description of the development process that is something more than listing out the activities, but something less than a step by step account. Specifically:

- What was the decision process and supporting development input that led to the attributes selected and their levels? What attributes were considered and discarded, and why?
- What assurances does the development process evidence give that respondents accurately interpreted the scale of the attribute changes? I ask this because, taken out of context on my first read, I had very little grasp of what the units on oysters and bass meant in the choice questions.
- How did the need to link attribute levels to models that can predict changes from the TMDL affect or constrain the choices of attributes and attribute levels?

3) Description of the data clear? Are conclusions drawn consistent with the data?

I found the description of characteristics of the sample, and the comparison of the sample to census and non-response summaries, to be informative and competently done. What is missing is a summary of responses to the choice questions, generally and conditioned on attribute levels. For example, the following would allow readers to gain a feel for the variability in the sample and the patterns of response:

- Proportion of sample that selected the status quo on all three questions
- Proportion of sample that selected a status quo and non-status quo among the three questions.
- Proportion of sample that selected non-status quo on all three questions.

I also think a conditional logit model with dummy variables for each attribute level (suitably normalized to sensible left out categories) would provide a good data summary. It would also allow the reader to gauge validity based on the comparative magnitude of dummy variable estimates. This is done to a degree in the sensitivity section, but I think it would be useful to present that here, as a data summary.

These summaries of the responses would provide a nice complement to the parametric models presented subsequently, and would help illustrate the variability leading to the main findings.

4) Data analysis

a) Are the specifications, variables, and methods consistent current state of literature?

In terms of estimation, the linear specification and logit errors are consistent with common practice, as is the use of a base model followed by richer inclusion of interactions. Consistent with my earlier comments, my one suggestion/concern here is to be clearer on the use of mixed versus conditional logit models. More specifically on methods, I would like to see discussion of what the random coefficient on the status quo is designed to provide, and how robust the results are to leaving this deterministic.

I have one request for clarification on the WTP calculations. In particular, it is not clear if the negative value for the status quo coefficient is used when computing the value for moving from status quo to some counterfactual level. I am assuming it is *not* used, since this would be consistent with avoiding hypothetical bias in the predictions. If it *is* used some addition predictions leaving it out are probably needed.

b) Do result sections provide adequate, clear, robust description of findings

In general yes, though as additional interactions are added to the model it becomes increasingly difficult

to get intuition on what is happening with marginal willingness to pay values. Some additional summaries surrounding table 13 would be useful, particularly as regards seeing the role of income effects in marginal WTP. Perhaps some figures showing point estimates and confident intervals for marginal WTP as income level changes? A similar strategy might be considered for table 23.

c) Interpretations and conclusions consistent?
Here too I would in general say yes, though the fact that the non-response analysis shows so clearly that the estimates are based on a selected sample makes me wonder what the consequences are for the parameter estimates of this. I don't have a specific suggestion, but this would be worth some thought.

On a different interpretation note, I would like to see some effort to place the household and aggregate WTP estimates in context by providing a comparison of the annual estimate to census figures for annual income in the study region. This would provide an additional margin for assessing plausibility and provide a more interpretable benchmark.

d) Non-response analysis appropriate and consistent with current practice?
This is a strength of the study. I particularly like how this effort allows the investigators to say something about the comparison in Likert scale attitudes, along with the usual difference in respondent characteristics. As noted above, the main question relates to how the non-response analysis is used, rather than how it is executed.

5) Sensitivity analysis
In general I found the robustness analysis to be well-executed. One place where some additional analysis could take place is in the area of repeated question robustness. Theory says that the repeated question approach is incentive incompatible, and so it would be useful to know if similar conclusions are drawn from a subsample consisting only of answers to the first question.

As mentioned a few times above, I am also unsure of what the random parameter on status quo is contributing. While it does show there is variability in the sample propensity to select a designed option regardless of attribute values, it also adds structure that should be sensitivity checked. How do the results change when the specification is simple conditional logit?

6) Total WTP
a) Approach to extrapolating household WTP to population consistent with best practice?
Choices needed to be made on how one does this, and I think the decisions made to extrapolate here are well described and reasoned.

b) Approach for estimating total WTP for users and non-users consistent with best practice?
It makes sense to me to divide estimates by the dummy for users versus non-users, so long as it is understood that these are still total values for each group, and that it is not possible to describe these as use versus non-use values.

I did have a few questions on interpretation. Specifically, the scenario described in table 16 is hard to interpret in the level. Are these 'big' changes or 'small' changes? Do the focus group and other

development work imply these are perceived as consequential changes? Some sense of this would help gauge the plausibility of the level WTP magnitudes, and (and mentioned above) their comparison to aggregate household income.

Also, in table 17 the point estimates jump out as potentially counterintuitive, in that out of watershed residents seem to garner higher benefits from the policy shock. Though these don't look to be statistically significant in their difference, I would argue that they are economically significant. Is there some further explanation for what is driving this, either mechanically (small differences in marginal utility of income?) or rationally?

7) Appendix

The appendix is complete and useful. I appreciated the opportunity to browse the detailed information on the focus groups and cognitive interviews.