Individuation of net benefits: Correct measure of aggregate net benefits is the sum across the population of individual net benefits, which involve a measure of individual costs and a measure of individual benefits. Individual benefits are the product of individual marginal WTP for a unit of risk reduction, times the number of units of risk reduction. If individual WTP per unit is independent of the sizes of the individual risk reductions, it is possible to calculate average WTP per unit and to multiply this by the aggregate risk reduction to get aggregate benefits. If there is a correlation between individual WTP per unit and the sizes of the risk reductions, however, the strategy of first averaging/aggregating, then multiplying will give a different answer from the correct strategy. This is true for the same reason that E[XY]=E[X]E[Y] only if X and Y are independent. **Implication**: For the "average/aggregate, then multiply" approach (typical of VSL-based calculations), it is necessary that individual WTP per unit and individual risk reductions be approximately independent. The decision to use this strategy requires that the analyst defend this (typically implicit) maintained hypothesis.

6. The white paper describes a simplified approach for updating the Agency's recommended mortality risk value estimate(s) (see section 5.1.1). This approach involves fitting a parametric distribution to the set of estimates from selected studies. This is similar to the approach used for EPA's current default VSL estimate.

a. Should EPA pursue this approach for updating its mortality risk valuation guidance in the near term (until a more detailed analysis can be conducted)?

It is important to drop the pretense that we are looking for one "true" number for WTP, which is the premise behind the current approach to combining the results from different studies.

This approach involves calculating a measure of central tendency for the marginal distribution of a variety of different WTP measures, representing a mix of different types of risk reductions and a variety of affected populations. These are demands for different goods by different groups of people. There can be no a priori assumption that these numbers measure the same thing, so the current process involves "apples and oranges." The constituent WTP measures differ for good reasons. They are not all measures of the same underlying fundamental constant. The goal should be to use WTP estimates that are matched as closely as possible to the policy context in question. For the EPA's needs, the types of risks which are most relevant are often not the case of "sudden death in the current period" that is the focus of wage-risk studies. Many health threats addressed by environmental policies consist of illness profiles with long latencies and substantial periods of morbidity prior to death. Lost life-years are not the sole determinant of WTP for health risk reductions.

As a practical matter, however, this requires that the inventory of WTP measures be sufficiently "thick" in the necessary domain. If the policy context represents a "gap" among available WTP estimates, there will have to be some sort of interpolation or extrapolation. Alternatively, placeholder estimates and sensitivity analysis may have to be used, but if this is done, it should be made *very clear* how the policy context differs from the contexts within which

the available WTP estimates have been measured. Only if these mismatches are clearly identified with researchers be able to identify where further research will be most useful.

b. If so, should the databases on which values are based be created using only one estimate drawn from each study or multiple estimates from each study?

The answer to this depends upon the reasons for different estimates in a study. Sometimes, these different point estimates are all, ostensibly, measures of the same thing that differ because of model uncertainty (or specification searching), or the use of different exclusion criteria. In this case, the research may report the sensitivity of WTP estimates to assumptions about the specification, but the researcher may ultimately identify the most-preferred model. In these cases, where one model clearly dominates others on a priori or statistical criteria, there is an argument for using only the "best" estimate from such a paper. However, across alternative specifications that cannot be rejected by the data, or across alternative exclusion restrictions where it is not obvious which set is most appropriate, it is essential to reflect the degree of "model uncertainty" in the estimates from this study, rather than to consider only the interval estimates of WTP stemming from just the preferred specification. Comprehensive error bars are needed, and these need to subsume both model uncertainty and parameter precision conditional on the choice of model.

In other cases, however, a paper will produce a variety of different point and interval estimates of WTP but this is because of heterogeneity in both the type of risk and the affected population. The white paper seems to imply that different estimates from the same paper can be used only if they are derived using separate models for separate samples. However, it is possible that the data can be pooled across separate subsamples with indicators for differences in the type of risk or the affected population, and one would *expect* different estimates. In such a case, the agency should choose those estimates from the study that correspond as closely as possible to the policy-relevant context, rather than averaging all estimates indiscriminately.

c. If only one estimate per study should be used, what criteria should the Agency apply in selecting the appropriate estimate? How would these criteria vary from one segment of the literature to the other? The paper describes the methods used to select independent estimates from each study. Does the Committee agree with the methods used?

The most appropriate estimate from a study is the one most closely aligned with the policy context, across studies of "sufficient quality." Sometimes, unfortunately, this will be no estimate. When the selection criteria cause the Agency to come up empty, it will be necessary to forge ahead with the closest available estimates, yet to acknowledge why they are inappropriate and to specify exactly what type of risk and what type of population constitute the relevant context. Any benefit-cost analysis should be qualified by the caveat that it was necessary to use a WTP estimate that was not really a good fit. As a practical matter, nothing more can be done until additional research has been completed concerning the appropriate context. But the Agency should persist in pointing out the theoretical ideal in each case, and the extent to which the available data fall short of this ideal.

The Agency needs to identify its "most wanted" archetypical WTP estimates, to set the research agenda for subsequent empirical work.

d. How important is it that estimates be drawn from non-overlapping subsamples? If multiple estimates per study are recommended in the construction of the meta-datasets, should the estimates be selected to avoid overlapping sub-samples?

Again, it depends upon the reasons for multiple estimates from the same study (or from different studies or different authors based on the same data sets). If one study ignores heterogeneity and estimates a central tendency only for some marginal distribution across the studied risk and population, and another study differentiates WTP estimates according to statistically significant heterogeneity, it is more appropriate to seek out the best fit of the study context to the policy context, rather than to focus on non-overlapping subsamples, which seems to presume that the identical underlying constant is being measured in all cases.

e. Does the Committee still favor analyzing the stated preference and hedonic wage estimates separately? If so, how should the separate results of these analyses be used in evaluating new policies? If not, how should they be combined in a single analysis?

Less important than the distinction between SP and RP studies is the distinction between illness profiles. Most RP wage-risk studies are limited to the illness profile consisting of "sudden death in the current period." Stated preference studies can be much more general. If the study scenarios have been appropriately designed, an SP study may also produce estimates of WTP to reduce the risk of sudden death in the current period. If the agency desired as estimate of WTP for this context, it is appropriate to use all information for WTP for the same type of risk reduction and the same affected populations. Revealed preference studies are less suited to illness profiles involving latency and long periods of morbidity in addition to premature mortality. It may be inappropriate to combine SP and RP studies in this case, not because of the differences in RP and SP data, but because the "good" for which WTP is being measured is fundamentally different (and possibly the affected population as well).

Even though many wage-risk RP studies may not address the same types of illness profiles as are relevant for the policy context of interest, these studies are still vitally important because validation of relevant SP studies may benchmarked against the "sudden death in the current period" illness profile when at least this profile is covered by both types of studies. Consistency between SP implications and best-practices RP studies, for comparable risks and populations (e.g. working age males) will be an important criterion for assessment of SP studies.

That said, it is still important to acknowledge that poorly designed SP studies will continue to be plagued by all of the potential problems with SP research that have been identified over the years, and many economists who have not been forced by necessity to resort to SP methods are still skeptical about SP as a class of research, rather than discriminating between better or worse SP evidence. Sometimes a weaker estimate of the right construct is will be greatly superior to a stronger estimate of the wrong construct.

The Agency can continue to encourage useful research by being very clear about the nature of the particular illness profiles that are most relevant to the policy context and being persistent in pointing out the deficiencies in the available WTP estimates.

There is likely some scope for using studies of risk-risk tradeoffs, despite the absence of WTP estimates, to translate RP estimates of WTP to reduce the risk of "sudden death in the current period" into WTP to reduce other types of risks. Of course, estimation errors would have to be compounded across these two stages.

As for expert elicitation, the committee acknowledges that the nature of the evidence in this context makes it less likely that individual experts will have significant private knowledge about WTP beyond that contained in available empirical studies. However, expert elicitation could be useful when it comes to the Agency's decisions about which estimates, for a specified risk and population, are a sufficiently good match and of sufficient quality to include in an analysis. The committee acknowledges that audiences will need to be carefully prepared to hear that the Agency will use different WTP estimates for different risks and different affected groups. Transparency in the process of how different studies are selected will be very important.

f. Would the Committee support the development and application of separate means or ranges generated from the two segments of the literature? Given separate means and/or ranges from each segment, should the results be weighted and combined to produce a single point estimate or range? If so, how? Are other presentations of the results preferable? More generally, how should uncertainty in the estimated value(s) of mortality risk reductions be handled in benefits analyses?

Using weighted averages of individual point estimates is only appropriate if these point estimates measure the same thing. Recent research certainly highlights extensive heterogeneity in WTP for risk reductions, both as a function of the type of risk to be reduced and the characteristics of the relevant affected population. If multiple estimates are available for the same context, then these can be averaged, and it is appropriate to consider some sort of weighting scheme that reflects the relative precision of the different point estimates.

The committee acknowledges that heterogeneity in WTP across types of risks will be more palatable to some audiences than heterogeneity across affected subpopulations. For example, the Agency has been taken to task for the "senior death discount." This is, however, a failure of communication with the general public, rather than any theoretical ambiguity about whether economics admits for different demands by different types of people. There is a clear need for some very accessible explanations as to why it is inappropriate to force people to bear higher costs for regulation than they would be willing to pay for themselves. It is true that it is somewhat difficult to convey the distinction between "the intrinsic value of different human beings" and the "different WTP of people in different circumstances." However, this difficulty does not constitute an excuse for using the wrong benefits measures for proposed policies. In the case of policies which represent unfunded mandates, there is an obvious need for a clear and

widely accessible (and regularly repeated) explanation for why economists seek to honor consumer sovereignty where possible.

Uncertainty in the estimated value(s) of mortality risk reductions should certainly be reflected in any benefit-cost analysis of policies. It is misleading to overstate the precision of the evidence. As mentioned above, model uncertainty is relevant, not just parameter uncertainty in the most-preferred specification. If multiple estimates of WTP happen to be available for the *same* type of risk reduction for the *same* population, then it is appropriate in any averaging process to acknowledge different degrees of precision associated with different estimates. Whether this is accomplish via Bayesian methods or by more conventional techniques is probably less important that whether it is done at all.