October, 1994

COMMENTS ON PROPOSED NOAA/DOI REGULATIONS ON NATURAL RESOURCE DAMAGE ASSESSMENT

U.S. Environmental Protection Agency
Washington, DC 20460

EXECUTIVE SUMMARY

EPA strongly supports the natural resource damage provisions of the Oil Pollution Act of 1990, the Clean Water Act, and the Comprehensive Environmental Response, Compensation, and Liability Act of 1980, as amended. We believe that contingent valuation (CV) is a useful methodology, particularly for determining passive use damages that cannot be measured in any other way. The practical choice is between using CV or implicitly assigning a zero value to passive use damages. We believe that CV, when carefully done, can provide reliable results for determining damages at a reasonable cost. Nothing in the remainder of these comments should be construed as compromising these basic EPA views.

EPA is very concerned, however, that the contingent valuation (CV) portions of the proposed NOAA/DOI regulations on natural resource damage assessment (NRDA) require unnecessarily expensive contingent valuation studies, and pose a great danger of freezing the development of the contingent valuation methodology. We present evidence documenting the bases for these concerns and suggest specific changes that will alleviate them.

We believe that allowing trustees the flexibility to use less expensive techniques for NRDAs is very important as long as the results can reasonably be relied upon by experts. It is important because the proposed regulations provide that conducting a CV study is the only way that passive use losses can be included in natural resource damages if they are to get a rebuttable presumption. In all likelihood, only a small number of the damage sites can justify the expenditure by trustees for NRDA that would be needed to satisfy the proposed regulations and guidance, even in those cases where the resources are later reimbursed out of the recoveries. Fortunately, we believe that there are many ways that NOAA/DOI can lower the cost of CV surveys, including eliminating the scope test as defined in the proposal, requiring more modest response rates, eliminating the preference for referendum format approaches, using lower cost survey techniques, and dropping the requirement for use of a survey research organization.

The available information indicates that passive use damages represent a very significant proportion of total natural resource damages (NRDs). Failure to include these damages in NRDAs would result in not adequately compensating the public for these damages and providing less than optimum incentives for those responsible for avoiding NRDs to avoid such damages. That, in turn, would make EPA's job more difficult because more pollution would occur than would otherwise be the case or than would be economically optimum. EPA believes that it is therefore very important that the regulations make every possible effort to include full and accurate passive use losses in NRDs. Unfortunately, we do not believe that the proposed regulations meet this objective. In general, the use of CV has been made unnecessarily expensive while the simplified procedures for determining NRDs entirely omit passive use damages. Specifically, we

suggest alternatives that emphasize accuracy and credibility, avoid introducing purposeful downward bias in CV results, remove the calibration factor, avoid dictating lump-sum payments, eliminate the preference for the referendum approach, avoid using screening factors, and modify the compensation formulas to include passive use damages.

We also suggest consideration of an alternative format that will avoid freezing the methodology that is used at the point that the regulations are written, even assuming that they represent the best methods available at that time. The regulations may also inhibit the search for new and improved methodologies since there would be no rebuttable presumption for assessments using improved methodologies for NRDA purposes. EPA advocates as flexible regulations as possible, avoiding unnecessarily prescriptive standards, and putting as many of the requirements as possible in the proposed guidance document. We strongly believe that much of the restrictive language concerning design standards which is contained in the Preamble should be removed.

In summary, EPA supports the use of CV to account for passive use natural resource damages. EPA also has an interest in using CV to carry out its own regulatory responsibilities in a flexible and responsible manner consistent with the different statutory mandates under which EPA operates.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

1. INTRODUCTION

These comments have been prepared in response to the *Federal Register* notices placed by National Oceanic and Atmospheric Administration (NOAA) on January 7, 1994[1] and by the Department of the Interior (DOI) on May 4, 1994[2] requesting comments on proposed regulations (hereafter the regulations) regarding either natural resource damage assessment (NRDA) in general (NOAA) or the use of contingent valuation (CV) in NRDA (DOI).

1.1. EPA EXPERIENCE IN DEVELOPING AND USING CONTINGENT VALUATION

1.1.1. Experience in Developing Contingent Valuation

EPA has contributed extensively to the development of the contingent valuation methodology beginning in 1973 and continuing through the present. During that period, EPA has provided extensive research assistance funding for both basic research on CV methodology and its application to determining the economic benefits of EPA programs. In the process, EPA research has resulted in more than 150 books, research reports, and articles on CV. A partial listing can be found in Appendix A.

1.1.2. Experience in Using Contingent Valuation

In recent years EPA has made increasing use of contingent valuation studies in determining the economic benefits of its pollution control activities. The determination of the magnitude of such benefits has been required since 1981 by various Executive Orders beginning with Executive Order 12291 issued in early 1981. In recent years EPA has increasingly turned to CV to determine the passive use (or non-use) benefits of proposed programs and regulations. Two recent examples are the draft Regulatory Impact Analysis on the Corrective Action Program and the Administration's proposed revisions to the Clean Water Act (USEPA, 1994). The regulation on which contingent valuation has played the most influential role has been the Navajo Generating Station regulation.[3] Although the CV study did not form part of the legal basis for the regulation, it played an important role in its formulation.

1.2. ORGANIZATION OF EPA COMMENTS

Because of EPA's particular concern with the contingent valuation aspects of the proposed regulations, most of our comments will concern the use of contingent valuation

[1] 59 FR 1062.

[2] 59 FR 23097.

[3] U.S. Environmental Protection Agency (1991).

in NRDA. Because of the more explicit statement of their proposed regulations by NOAA, their proposed regulations will be used as the basis for these comments. Although the corresponding sections in the DOI *Federal Register* notices are not usually listed, our comments are in most cases equally applicable to the DOI proposals. All page references to the NOAA/DOI proposals will be to the NOAA January 7, 1994 *Federal Register* notice.

Section 2 will discuss reducing the cost of carrying out CV studies. Section 3 concerns making it easier for trustees to include full and accurate passive use damages. Section 4 will discuss how the NOAA/DOI regulations can reduce the risk that they will freeze further advances in CV methodology and provides some more general observations.

2. REDUCE THE COST OF CONTINGENT VALUATION STUDIES

2.1. WHY REDUCING THE COST IS VERY IMPORTANT

One common thread that flows through many of our comments concerns the question of the cost of doing contingent valuation studies as proposed by NOAA/DOI. The costs should be kept as low as possible consistent with assuring adequate reliability so as to make available this important and vital information in as many cases as is cost effective. If the cost is more than modest, very few CV studies will be carried out. Since the great bulk of the damages are likely to be in middle-sized cases, this would be a substantial loss to the trustees. It would also greatly decrease the economic incentives for polluters to be careful.

2.2. USE COST-EFFECTIVE APPROACH TO SELECTING CV REQUIREMENTS

The gains from elaborate data gathering protocols must always be balanced against the cost of those protocols. Thus, before any technical requirement is added to the NRDA process, four questions should be explicitly asked and answered:

(1) Does the requirement improve the reliability of the results?
(2) Is the improvement in reliability provided by the requirement appropriate to allow experts to rely reasonably on the results of the assessment?
(3) Is the cost of the requirement justified by the increase in reliability?
(4) Could a sufficient improvement in reliability be achieved at less cost?

We believe that NOAA/DOI should answer each of these questions for each of the cost-increasing requirements included in their regulations and preferences stated in their preamble. Unless the answers to questions (1) through (3) are "yes," and the answer to (4) is "no," we do not believe that any requirement should be included. We believe that all of the requirements discussed in the remainder of this Section do not meet this test.

Obviously, there is a trade-off between requirements to increase reliability and the cost of these requirements. We recommend a balancing of the benefits and costs of expected changes in reliability. The issue should not be what requirements can be added that might conceivably improve the reliability of the studies without regard to the cost, but rather what are the benefits of each proposed strengthening of the requirements and what are the costs of them? In general, we are concerned that there is insufficient attention both to the cost aspects and the expected improvements in reliability of the regulations being discussed.

Many of the proposed requirements and recommendations in the proposed regulations for conducting CV studies for NRDA would impose considerably higher costs on the trustees than might be required if other available approaches were used. In many instances there is not clear evidence that these recommended, and more costly, approaches will result in more accurate CV results. Higher costs per se are not a measure of higher reliability.

Examples of such requirements/recommendations that would increase the costs of CV studies to trustees without a demonstrated increase in reliability relative to lower costs alternatives include the following:

2.3. Split sample, three version, two-way scope tests
2.4. Referendum question formats
2.5. Minimum 70 percent response rate
2.6. In-person interviews
2.7. Use of survey research organization

Each of these will be discussed in the remaining subsections in this Section.

NOAA/DOI have not given adequate consideration to cost-effectiveness for CV, although NOAA acknowledges that this is an important consideration for trustees in other areas such as that of benefits transfer. We find it strange that NOAA/DOI recommend consideration of the replacement cost method, which is not a measure of the value of lost services, but does not consider less costly CV approaches. High costs do not guarantee good quality work. They guarantee only that many trustees will do without useful information that they might have been able to obtain by using less costly techniques. No convincing demonstration has been made that these more expensive techniques would lead to any additional precision or other benefits to justify the additional costs. The clear implication of the proposed regulations, unfortunately, is just the opposite: that the less costly techniques are not capable of providing useful or valid information. This implicit conclusion is costly and damaging for trustees and others such as EPA responsible for protection and management of natural resources who need to use CV to carry out their responsibilities.

2.3. ELIMINATE SCOPE TEST AS NOW DEFINED

Because of the limited evidence available on the cost and implications of the proposed scope test, EPA supported a study of some the theoretical, statistical, and methodological issues involved. The draft reports resulting from this study, conducted by the University of Colorado, are included as Appendices B and C. Appendix D presents some comments on the proposed NOAA/DOI scope test for CV by Professors Kenneth Arrow of Stanford University, Edward Leamer of the University of California at Los Angeles, Howard Schuman of the University of Michigan, and Robert Solow of the Massachusetts Institute of Technology. All were members of the NOAA "Blue Ribbon Panel." Professors Arrow and Solow are recipients of Nobel Prizes in Economics. In the comments reproduced in Appendix D, they state that there is a very sharp conflict between the basic character of the proposed NOAA/DOI scope test and the sense of the NOAA panel. They further say that "We fear that the proposed test will increase the cost of the surveys with no compensating increase in their 'reliability.'" They conclude that:

> The fundamental problem with the statistical definition of sensitivity is that it depends ... on the sample size. In small samples, no effects are "statistically significant." In large samples, everything is "statistically significant." What that means is that the proposed scope test can almost surely be passed if the trustees are willing to pay a high enough cost. But the willingness to bear this cost has no obvious implications for the "reliability" of the results.

This appears to be a very basic problem with the NOAA / DOI proposed scope test. Until and unless it can be resolved, we favor dropping the scope test as currently defined. If for any reason this advice is not followed, we have some more detailed comments and suggestions on the proposed test, which will be discussed in the next three subsections.

2.3.1. Use Within Rather Than Between Subject Performance Tests

The proposed NOAA / DOI regulations appear to be based on the premise that absolute values (*i.e*, between subject differences) are unbiased and that relative values (*i.e*., within subject differences) are biased, perhaps due to sequence effects. Based on *a priori* theoretical arguments discussed in Appendix B, between respondent comparisons are shown to be statistically less powerful (and hence require larger sample sizes) than within respondent comparisons (see Erlebacher, 1977, Rosenthal and Rubin, 1980, and Keren, 1993). Similarly, evidence from psychology suggests that absolute judgments obtained from a split sample, as suggested in the proposed regulations, are not necessarily better than relative judgments (Baird and Noma, 1979). On the other hand, within sample tests are subject to sequence bias, which might be present when respondents are asked for more than one value (Rosenthal and Rubin, 1980). This issue is addressed in Appendix B as such an effect, if present, might negate the statistical benefits.

A CV experiment was conducted which showed that sequence effects do not significantly bias successive bids for different scenarios which vary in scope. However, as shown in Appendix B, the requirement of using a split sample for conducting the proposed scope test comes at the cost of an increased sample size necessary to show a statistically meaningful difference in scenarios. This increase in sample size is on the order of at least a factor of 4.2. The research reported in Appendix C illustrates that the implementation of the proposed scope test procedures proposed by NOAA could require much larger sample sizes. In the absence of additional research or other evidence to the contrary, EPA concludes that the split sample requirement is unnecessary, excessively costly, and based on speculation rather than science.

A comparison will be made in this and three other cases later in this Section with lower cost alternatives where no convincing evidence exists in the refereed literature of lower reliability for the lower cost alternative. The increase in costs will be expressed in each case for which data exists as a lower bound, a best available conservative estimate, and an upper bound.

A split sample scope test increases sample size (and therefore survey costs) compared to a within subject test by at least a factor of two because twice as many respondents are needed. However, as discussed above, and as Schulze and McClelland (1994) in Appendix B demonstrate theoretically, the factor is actually higher due to the lower power of between sample scope tests relative to within sample scope tests. In their empirical example, the increase in sample size needed was 4.2, so that the ratio of the cost of a split sample test to a within subject test would be 4.2 to 1 (hereafter referred to as a cost ratio factor, cost factor, or simply factor of 4.2). The lower bound of the increase in cost is then a factor of two. The best available estimate is a factor of 4.2. No upper bound estimate is available, but is presumably larger.

2.3.2. Eliminate Increases in Scope from Performance Tests

It should also be noted that the proposed scope test requires three options so that both an increase and a decrease in scope can be shown to differ significantly from the base option. The research presented in Appendix C suggests that a serious problem exists with this proposal. Using market demand data it is shown that, because demand is downward sloping, it is difficult to show a statistically significant increase in WTP for an increase in scope (quantity) from average current consumption. Diminishing marginal WTP implies that the effect of increases in scope on WTP will be much smaller and more difficult to prove than will the changes in WTP resulting from decreases in scope. Further, without *a priori* knowledge of the position of the demand curve, choosing the three levels of scope so that changes in scope provide statistically significant changes in WTP, while satisfying the constraint limiting the magnitude of scope changes, is a difficult task (given diminishing marginal WTP and the inherently large variance in public good values). The impact on study costs of performing two rather than one scope test is first to increase the sample size required by a factor of 1.5. However the actual cost increase is likely to be much larger since the sample size must be further increased to detect the likely small incremental value of an increase in scope.

Thus the lower bound of the cost increase from this requirement relative to a one-way scope test is a factor of 1.5. The best available estimate is greater, and the upper bound would be still larger.

2.3.3. Eliminate the 95 Percent Rule

It is important to avoid setting standards for which careful research does not yet exist or where it is likely that further research will overturn the basis on which the standard was set. A good example is provided by the so-called “95 percent rule” included in the NOAA proposed regulation. As discussed above, the proposed regulations require between subject tests without adequate investigation of the costs and benefits. In addition, they require a preliminary within subjects test which must equal or exceed 95 per cent of the subjects. Since NOAA has been unable to explain how the 95 percent was determined, it appears

likely that it amounts to nothing more than a guess by the authors. Future research may or may not support the wisdom of the 95 percent figure, but its use in a semi-permanent regulation is premature and ill-advised.

2.4. ELIMINATE PREFERENCE FOR REFERENDUM FORMAT APPROACHES

CV question format refers to the exact formulation of the question asked of respondents from which estimates of WTP are based. Two elements of this question format are the choice mechanism and the payment vehicle. The choice mechanism refers to the way the valuation question is asked, *e.g.*, saying yes or no to a specified dollar amount, picking a dollar amount from a list of amounts, giving a dollar amount in response to an open-ended question, or stating a preference for one alternative or another that each involves a cost and a level of benefit. The payment vehicle refers to the how respondents are told they would be paying, e.g., through higher taxes, higher prices for gasoline, higher utility bills.

NOAA states a strong preference for the referendum format, but fails to present sufficient empirical evidence to support this stated preference (page 1144). Available evidence does not support the conclusion that this format is necessarily preferable for meeting the stated general goals. Although the referendum format is a useful tool, it is just one of many possible question formats, and should not be singled out as the preferred format at this time. It has strengths and weaknesses, as do many other formats that could be considered. The NOAA/DOI selection of the referendum format is unsupported and insupportable based on the available refereed literature. EPA recommends that NOAA/DOI state the general goals for question format design and remove the stated preference for the referendum format.

Many different choice mechanisms and payment vehicles have been used in CV studies and new variations continue to be developed. All question formats used to date have some strengths and some weaknesses. Some, such as iterative bidding questions, are not used very often because they have been shown to introduce biases in the responses. Others, such as open-ended questions, present difficulties because they tend to elicit a higher rate of non-response. Still others, such as referendum question formats, present questions that are believed to be easier to answer, but that do not provide as much information to the researcher, thus increasing sample size requirements and the complexity of the interpretation of the responses. Research on CV techniques continues to evolve and develop new ways to address the limitations that have surfaced for each of the types of question formats that have been developed. There is not sufficient evidence available at this time to support the position taken by NOAA that the referendum format produces the best results.

On page 1144, NOAA recommends that "trustee(s) use a choice mechanism and payment vehicle that are both credible and incentive compatible, *i.e.*, do not impose strategic bias." These goals can be further generalized to two basic themes. The question format should be:

- *Realistic and easily understood.* Because the question is hypothetical it is very important that the respondent understand the question and find it realistic that he/she be called upon to make such a judgment.

- *Neutral and unbiased*, such that respondents are not influenced by the question to answer with anything other than their best estimate of their true WTP. Biases can result from information provided in the question that influences the respondents (*e.g.*, starting point bias) or from a perceived opportunity to influence the results of the study by providing an inaccurate answer (*e.g.*, strategic bias).

One basis for the NOAA endorsement of the referendum format is that it poses a voting decision context, which NOAA argues is a familiar context for the public in decisions regarding provision of public goods. It is correct that a voting decision is a reasonable and realistic context for decisions regarding public goods and is therefore a context that should be considered in designing a CV instrument for NRDA cases. However, a yes-no question is not the only way to use the voting context. A yes-no question might be, "Would you vote yes or no on a proposal to do ... if the cost to you would be $X." Another way to use the voting context might be, "Suppose there is to be a vote on ... What is the maximum amount you will be willing to pay and still vote yes?"

There may be other contexts that are also appropriate to consider. For example, prevention of oil spills may involve actions that would increase the price of fuels. In this case a context of higher prices for fuels might also be a realistic context for the CV questions. A voting context is one of several realistic options to consider, not the only one worth considering. What needs to be stressed here are the goals of realism, ease of comprehension, and similarity to familiar decisions. There are many ways that these goals can be achieved.

An important limitation of the yes-no referendum format that has bearing on the question of whether the information obtained is better than when other formats are used is that the statistical analysis required to derive mean WTP estimates from the yes-no responses is much more complex than what is required when maximum WTP estimates are directly obtained, such as with payment cards or open-ended questions. One important implication of this is that the sample size required for statistically adequate mean estimates is much higher, which results in significantly higher survey implementation costs, especially for in-person interviews (see Cameron and James, 1987, and Appendix B). Appendix B demonstrates in a CV experiment that the sample size necessary to pass a one-way scope test is increased by a factor of 3.3 by use of the referendum approach in comparison to open-ended WTP. This is very close to Kanninen's (1993) analysis that indicates that four times as many single-bounded dichotomous choice CV observations are needed to achieve the same level of efficiency as can be achieved by using an open-ended

CV format.[1] Another implication is that the derivation of the mean WTP estimates may be sensitive to survey design factors. For example, Duffield and Patterson (1991) found that the estimated parameters were sensitive to the allocation of the sample across the payment amounts while holding total sample size constant. In a similar vein, Boyle and Bishop (1988) found the results sensitive to the specification of the logit model used to estimate mean WTP from the referendum responses.

The best available estimate is that the referendum approach increases survey costs by a factor of 3.3 compared to open-ended willingness to pay (WTP). The lower and upper bounds are unknown, but are presumably smaller and larger than 3.3, respectively. If the cost of a survey using open-ended WTP is $50,000, then the best available estimate for the referendum approach is $165,000.

A second important reason to drop the preference for using the referendum format approach is that there is some evidence that its use elicits WTP estimates that are higher than those obtained using open-ended questions. This aspect will be discussed in Section 3.7 below.

In summary, this subsection (and Section 3.7 below) demonstrates that use of the referendum approach is not a conservative methodology in that, when compared to open ended WTP questions, the referendum approach consistently provides larger values. Further, experimental economics research shows that a similar institution, the posted offer market, provides upwardly biased values in early rounds which are inherently similar to those provided by a "one shot" CV study. This concern is greatly exacerbated by the demonstration in Appendix B that the referendum approach is statistically inefficient in that, to show a statistically significant difference between the options evaluated there, it required a further increase in sample size of a factor of at least 3.3 over that needed in a study using open ended WTP. For these reasons, we recommend that NOAA/DOI eliminate the current preference for the referendum format. Future research may or may not provide the basis for such a preference, but current research does not; in fact, it appears to suggest the opposite. Until more adequate research is available, it appears prudent to avoid a preference in the regulations for any particular survey format such as the referendum format.

2.5. REQUIRE MORE MODEST RESPONSE RATES

With respect to the response rate issue, we believe that there are increasing marginal costs to achieving higher response rates. Because of the importance we place on keeping costs as low as possible given the need to insure reliability, this trade-off is a very important issue and should be made on the basis of careful research on the relationship between

[1] Table 2, p. 144.

increased reliability and increased cost, not what appears to be an *ad hoc* guess. Until careful research is available on this issue, we recommend omitting any required response rate, and certainly one as costly to achieve as 70 percent.

The objective of a high response rate is to reduce the existence and magnitude of any potential response bias introduced by systematic differences in values held by respondents as opposed to non-respondents. NOAA is correct when it states, in response to comments, that "there is no bright line to determine at what level of response a survey's results become unreliable,"[2] but NOAA in effect has established just such a "bright line" at 70 percent. The literature does not substantiate that a bright line should be established at a 70 percent response rate. A 70 percent response rate, while achievable for all survey modes, can greatly increase expenses compared to slightly lower rates, sometimes with little extra benefit. A 70 percent response rate is perhaps the absolute maximum response rate that one can achieve with telephone interviews using random digit dialing and using intensive refusal conversion methods. In essence, this criteria is likely to implicitly exclude telephone survey approaches.[3]

It is standard survey practice to evaluate and account for differences between respondents and non-respondents using follow-ups with non-respondents, and using comparisons of characteristics of respondents to the population and to non-respondents, if available (Water Resources Council, 1983; Schulze *et al.*, 1993; Rowe *et al.*, 1992; Loomis, 1987). Through these techniques, a survey with a 65 percent response rate, which would not pass NOAA's "bright line," would provide results that are virtually equally defensible to the same survey with a 70 percent response rate.

Also, response rates should not be used as a basis for preferring in-person surveys to mail surveys. Mail surveys can achieve 60 to 80 percent response rates through well-designed surveys and through incentives for completion (James and Bolstein, 1990; Dillman, 1991; Doyle *et al.*, 1991; Rowe *et al.*, 1993; Schulze *et al.*, 1993). In well-executed comparison studies, in-person response rates seldom exceed mail response rates by more than a few percent where both approaches use repeated attempted contacts (Goyder, 1985). In a recent study comparing mail and in-person surveys in the Detroit area, both approaches had very similar response rates and results for all response groups except the in-person surveys had higher response rates for inner city residents (Krysan *et al.*, forthcoming). Many government studies require human subject rules that require that the

[2] Page 1162 column 1.

[3] Without incentives, mail surveys also are unlikely to reach 70 percent completion rates. For example, the most recent U.S. Census exceeded 70 percent by mail only through the last iteration where the envelope was marked "response is mandatory." (Dillman *et al.*, 1994).

respondent be informed that a survey is voluntary and require the in-person interviewer to accept a refusal. As a result, in-person response rates seldom exceed 70 to 80 percent. The exception is extremely well funded government studies where the government sponsorship is identified (Heberlein and Baumgartner, 1978), but explicitly listing sponsorship is an issue of debate for CVM studies applied to natural resource damage assessments.

EPA recommends that NOAA focus on establishing the objectives of high response rates and detailed evaluation of, and adjustments for, non-response based on characteristics of respondents and non-respondents. These efforts can be accomplished through small follow-up surveys and through comparisons of population data to respondent data.

NOAA should indicate that a range of 65 to 75 percent is consistent with high quality respondent-friendly methods that can be implemented by in-person, mail, and sometimes with phone surveys. NOAA should avoid setting any specific "bright line." A 65 percent response rate can be more routinely achieved with reasonable costs while losing little in accuracy that cannot be accounted for in follow-up analyses. NOAA can suggest that reliability is expected to increase with increased response rates.

NOAA indicates that "The trustee(s) shall document the rationale for the selected response rate." [§990.78 (a)(5)(ii)(A)(5)] The meaning of this clause is not clear, particularly given that the previous clause mandates at least a 70 percent response rate. It would not seem to be consistent with the requirements of "reasonable cost" (in the prior subsection) to obtain response rates much above 70 percent given the significantly increasing expenses associated with obtaining higher response rates, especially for in-person surveys.

2.6. USE LOWER COST SURVEY TECHNIQUES

There is no research we are aware of to support the unconditional superiority of in-person or several other types of surveys in all circumstances. In-person interviews have been shown to be open to interviewer bias as well as biases due to social desirability or compliance issues. The substantially higher costs involved in using in-person interviews are difficult to justify for intermediate size natural resource damages even though value estimates are often crucial to obtaining substantial recoveries. Appendix D provides the views of the Chief Survey Methodologist of the Census Bureau on this subject. Trustees should be encouraged to carry out high quality, cost-effective studies that may require one or a combination of survey techniques depending on the circumstances and the cost of each in those circumstances. More expensive studies are not necessarily more accurate studies.

The proposed regulations require that "The trustee(s) shall document the rationale for the selected mode of survey administration." [§990.78 (a)(5)(ii)(B)(1)] This allows all survey modes to be selected and defended on equal grounds, which EPA feels is appropriate, especially if NOAA makes clear the objectives to be meet. In the preamble, however, NOAA continues to focus on in-person survey methods rather than survey administration

objectives. The language that "... trustees seriously consider the use of in-person interviews for the final survey..."[4] and "NOAA anticipates that this documentation would include a discussion of the factors that led the trustee(s) to reject use of in-person interviews."[5] is inconsistent with the regulatory language and has the implicit force of virtually requiring in-person surveys.[6]

The cost of survey administration is an important consideration. Throughout the NRDA regulations, NOAA (and DOI) requires assessment to be at reasonable costs. For example, NOAA states that research should be designed to be "consistent with the requirements of reasonable costs in order to ensure reliable inferences to the general population."[7]

The relative costs of different survey modes varies considerably depending on the number of follow-ups, the dispersion of the population to be interviewed, incentives provided, the length of the survey, and other factors. Typical incremental costs for mail and phone CV type surveys are between $20 to $40 per completed survey. Typical incremental costs for in-person CV type surveys are between $200 and $500 per completed survey.[8] The lower bound of the cost increase is thus a factor of five, while the upper bound is 25.[9] One way to determine a best available estimate would be to take the ratio of the average of the upper and lower bounds of the cost of in-person surveys relative to the average of the cost of mail surveys, which yields a factor of 11.67.[10] These cost differences are so substantial that they may inappropriately preclude many trustees from being able to even consider undertaking CV studies, and may therefore preclude many trustees, and the public, from receiving appropriate compensable damages.

---

[4] Preamble page 1144, column 3 at line 40 and page 1162, column 2 at line 66.

[5] Page 1145 column 1 at line 9.

[6] The Preamble language is important because "NOAA is proposing that reliable estimates of passive use value due to discharges of oil can be estimated using CV so long as the CV study follows the guidance offered in this preamble and the proposed regulations." [Page 1074, column 2 at Line 4]

[7] Preamble page 1144 column 2 at line 2.

[8] We know of no estimates in the CV literature for these cost estimates. They are based on estimates by practitioners with knowledge of prevailing costs.

[9] $200 divided by $40 yields a factor of five; $500 divided by $20 yields 25.

[10] An average of the upper and lower bounds of the costs of mail surveys is $30 per interview. The average of the upper and lower bounds of the cost of in-person interviews is $350 per interview. This yields a ratio of 11.67 to 1 for in-person compared to mail.

NOAA may suggest that the added costs of in-person surveys are considered, by NOAA, to be acceptable if this survey mode is selected by a trustee(s). But EPA contends that NOAA should not explicitly or implicitly require a substantially more expensive survey mode without certified, verifiable, and substantial improvements in the reliability and validity of survey results obtained with the more expensive method. EPA believes that there is insufficient evidence to routinely pass this hurdle.

In the remaining subsections of Section 2.7 we address the specific reasons that NOAA asserts for giving generic preference to, and virtually mandating, in-person surveys. EPA finds that these reasons are either not compelling or sufficiently generic in nature to uniformly prefer in-person surveys, especially given the substantially higher costs of in-person surveys. **EPA recommends that no survey mode be singled out as unequivocally better in either the regulations or the preamble. Rather, the regulations should call for using the mode most appropriate to the survey being undertaken and implementing the survey using the best methods available for that mode.**

2.6.1. Sample Frame Coverage Rates

NOAA generically states that probability sampling is exceedingly difficult with mail surveys[11] as one reason to support the use of in-person surveys. This issue is tied to concerns about complete coverage of the target population in the sample frame. This concern is in some cases legitimate if nationwide mailing lists are utilized. However, the assumption that such lists will be used to construct the sample for all CV studies is incorrect. NOAA has ignored (or left unstated) several factors, including:

- In-person interviews can suffer from incomplete coverage problems because of restricted access, unknown locations, and other limitations.

- Mail surveys can be targeted to addresses, rather than names, and then select individuals within the household to respond, which can often replicate the in-person process with as high coverage and response rate. Random digit dialing phone surveys can also achieve high coverage rates and can be used to make initial contact for a mail survey.

- While non-coverage in the sample frame may increase uncertainty, evidence does not suggest that it substantially biases the results obtained in mail surveys. Evidence from Thornberry and Massey (1987) suggests that in the mid-1980's only seven percent of households did not have telephones, and these households are more likely to have unemployed heads of household and lower incomes.

[11] Preamble page 1144, column 2 at line 63.

More important is the potential exclusion of households with non-listed telephone numbers. Unlisted phones may be held by people with low incomes or with high incomes, and are increasingly held by females. Conventional wisdom is that unlisted phones run across the entire social and economic spectrum (Lepkowski, 1987). As reported in Chestnut and Rowe (1990), unpublished comparisons of phone survey results using listed phones and random digit dialing at Washington State University Social and Economic Survey Research Center found little difference in the average attitudes across those with listed and unlisted phones. While there may be some differences in the average attitude of households with listed telephones versus households with unlisted telephones, available information does not suggest that non-coverage bias due to omitting these unlisted phones is substantial or to infer whether the overall impact of such bias would increase or decrease population average WTP results in CV surveys. On this basis, adding millions to the cost of survey data collection must be questioned as being cost effective. Further, as noted above, unlisted households can be sampled by use of random digit dialing to make initial contact either for a mail or a telephone survey.

- For many CV surveys, adequate mailing lists are readily available, *e.g.*, people with hunting or fishing licenses.

EPA recommends that NOAA focus on the objective of high coverage rates plus the evaluation and correction of potential non-coverage bias, rather than establishing an explicit generic preference for in-person surveys.

2.6.2. Self Selection Biases and Response Rates

NOAA suggests that in-person surveys are preferred to mail or phone surveys because (1) in-person surveys make it more difficult for respondents to self-select whether they will participate in the survey,[12] thus reducing potential self-selection bias, and (2) will result in higher response rates and (lower item non-response rates).[13] These concerns have been addressed in mail surveys through high response rates, careful survey design, and follow-up analyses.

CV studies by mail do allow respondents to review the survey and decide whether or not to participate. Evidence suggests that respondents who respond later in the process (perhaps after multiple contacts) may have lower values than those who respond quickly, which may imply that non-respondents may have still lower values. But, evidence also suggests that the degradation in values is not necessarily so large as to dramatically affect

[12] Preamble page 1144, column 2 at line 68.

[13] Preamble page 1144, column 3 at line 71.

the results. For example, Schulze *et al.* find that average WTP for cleaning up a hazardous waste site in Montana was not significantly different when comparing respondents who returned their mail survey after the initial mailing to respondents who returned their the second and third mailings, although average values decreased slightly. These types of influences are routinely accounted for in CV assessments through comparisons of respondents to the population and, if possible, to a sample of non-respondents.

Mail (and in some cases phone) surveys can obtain very high response rates, and very low item non-response rates,[14] and already will be required to do so under the NOAA regulations, which reduce the potential significance of self-selection bias. For evidence of high response rates in mail surveys, see Goyder (1985), Heberlein and Baumgartner (1978), and Dillman (1991), and Doyle *et al.* (1991).

EPA suggests that NOAA require trustees to minimize, evaluate, and as possible correct for any potential self-selection bias as a more cost effective requirement than using this concern as a criteria to generically support in-person surveys, especially given the requirements of high response rates.

2.6.3. Complicated Instruments

NOAA correctly indicates that complicated survey instruments requiring many branches and extensive audio or visual materials may be better implemented in-person. But, it is not at all necessary to give uniform preference to in-person CV surveys on this basis because:

- The information in any CV survey (and most all surveys) must be presented in a manner and length that is accessible to the general public in the sample frame. If the information is so complicated or extensive, the success of a survey will be hampered regardless of the survey mode. The researcher may be better advised to address portions of the problem with different instruments and respondent samples.

- Complicated branching can be readily addressed in both phone, mail/phone, and in-person interviews (Dillman, 1978). Complicated branching often can (but not always) be avoided in mail surveys by using multiple survey versions, easy skip

[14] For example, Rowe *et al.* (1992) obtained responses rates in excess of 70 percent on a mail survey addressing natural resource damages from an oil spill in the Pacific Northwest. A recent mail CVM for a NRDA in Montana achieved response rates of 68 percent (75 percent including telephone follow-ups) and item non-response rates of less than one percent for all questions except income and a WTP follow-up question, which had non-response rates of less than eight percent (Schulze *et al.*, 1993).

patterns, and increased sample sizes—all at a cost much less than in-person interviews.

- Many visual materials can be provided in mail surveys (or as mail-outs supporting phone surveys) at significantly reduced costs compared to in-person interviews.

NOAA should not uniformly recommend in-person interviews on this criteria, but rather suggest that when complicated branching, or extensive visual materials are required, this may be cause for trustees to consider in-person survey mode.

2.6.4. Within Household Random Selection

NOAA indicates that the selection of an individual within a household is difficult to assure with mail surveys and that in some cases more than one person may have input to the survey. NOAA is placing too much emphasis on this consideration because:

- Random selection is often difficult to assure with in-person and phone surveys because the selection process often results in reduced response rates.
- There are strategies to select individuals within a household on mail surveys. For example, the cover letter can specify which individual should complete the survey (female head of household, male head of household).
- It is not clear that input of either (or multiple) heads of households is inappropriate in a CV survey. In most total value CV studies, the value question address household WTP and values are aggregated across households, not individuals. If multiple household heads confer on responses we expect a more carefully considered response that better reflects the household values of interest.
- In some cases, survey response data can be compared to the mailing list data, by survey ID, to determine whether such sampling has been followed for each survey mailed. To the degree that differences exist such as females are over sampled, data analysis of survey responses can be used to examine the potential significance of any bias introduced and to correct for this potential bias.

Given these considerations, the within household selection criteria is overemphasized as a justification to prefer an expensive in-person survey mode for CV applications.

2.6.5. Holding the Respondents' Interest

NOAA suggests that the presence of an interviewer will motivate and hold the respondent interest and thereby result in improved responses.[15] An in-person interview

[15] Preamble page 1144, column 3 at line 53.

mode does not make the interview more interesting; the survey instrument quality is a separate issue. It is true that an interviewer may help to reduce distractions. If NOAA is asserting that responses to mail surveys are less carefully considered, NOAA should provide evidence to support these claims before removing trustees' flexibility to use alternate methods. In fact, NOAA acknowledges that commenters have suggested, and EPA concurs, that mail surveys allow respondents the opportunity to consider their responses, rather than provide immediate responses, to sometimes difficult CV questions. This feature is an advantage of mail surveys.

There are equally important and well documented offsetting concerns with in-person surveys (Dillman, 1978; Babbie, 1973; and Bailly, 1978). These concerns include the potential for: respondents feeling the necessity (responding to NOAA's interviewer motivation) to complete the instrument regardless of their interest, which may result in inaccurate results; respondents providing answers they perceive are desired by the interviewer, which introduces social desirability bias; and the desire to complete the survey immediately, which may preclude the ability to spend time to give more carefully considered responses.

Given the requirement for high response rates and lack of evidence to suggest improved responses to CV surveys through interviewer motivation, EPA suggests that this criteria should not be used to generally prefer in-person surveys.

2.6.6. Control of Question Order

NOAA suggests that respondents to mail surveys may read ahead,[16] which could affect the results. This is correct, but there is no evidence that this is a problem for mail surveys.

This argument overlooks a substantial literature in the 1980's beginning with the work by Schuman and Presser (1981), that shows that face-to-face and telephone questionnaires often produce large order effects; *i.e.*, the order of the questions can influence the responses. Two recent studies addressing this question for mail surveys (Bishop *et al.*, 1988 and Ayidiya and McClendon, 1990) have show that self-administered surveys are somewhat less subject to these order effects than are face-to-face interviews. Rather than being a negative of the mail survey approach, this may in fact be a relative strength, vis-a-vis in-person interviews because there is less ability for question order to manipulate results.

Mail surveys are designed recognizing that respondents may read through them and therefore do not have “tricks” in them. Further, if a respondent reads ahead they may better understand the issue about which they are responding and therefore may provide better, not worse, answers to all questions. Since mail surveys typically place information in easy-to-answer questions at very short intervals to keep respondents involved, it is relatively

[16] No control of question sequencing (Preamble page 1144, column 2 at line 70).

easy to determine if material has been skipped by examining item non-response rates. These are typically very low, especially for informative material. EPA concludes that this criterion should not be used to support in-person surveys.

2.6.7. Recording of Respondent Responses to Open-Ended Questions

NOAA suggests that in-person surveys allow the interviewer to record, verbatim, responses to open-ended questions. Typically, after the pretesting phase, CV surveys do not require many open-ended questions. Mail surveys allow respondents to write their desired responses, and to even review and edit their responses, which can be transcribed verbatim. Phone surveys also allow interviewers to record, verbatim, respondent answers to open-ended questions. These factors provide little support for the stated preference for in-person surveys in the Preamble.

2.6.8. Types and Applications of In-Person Surveys

While supporting in-person interviews, NOAA has failed to differentiate between different types of in-person surveys and their relative effectiveness for CV applications. Are all in-person surveys to be given equal preference regardless of whether they are conducted in the home, as mall intercepts, as on-the-street intercepts, or through group interviews at civic functions (Church, rotary clubs, etc.)? EPA suggests that many in-person surveys strategies are inferior to other survey modes for CV applications. CV surveys are sufficiently complex, usually require supporting materials that must be carefully considered, and are to reflect one household's opinion rather than a group of household's opinions. EPA suggests that NOAA make clear the types of in-person surveys it considers to be most defensible.

2.7. DROP REQUIREMENT FOR USE OF SURVEY RESEARCH ORGANIZATION

NOAA is correct, in our view, in proposed §990.78 (a)(5)(ii)(B)(2) in requiring that trained and supervised interviewers be used for in-person interviews. But, NOAA should not *mandate* the use of professional survey organizations for all CV surveys [proposed §990.78 (a)(5)(ii)(B)(3)]. While it may be desirable to use such organizations, this is not necessary in all cases. For example, for mail surveys a survey organization is less critical as experienced individuals can direct such efforts regardless of whether they are in a formal survey organization.

NOAA should advise any trustee, in supplemental guidance documents but not in the regulations or preamble, that to defend a CV survey may require an individual or organization with substantial experience in survey administration. The requirement of a professional organization is not necessary. Further, NOAA is not clear what constitutes a "professional survey organization." Does this include universities with survey units? It should. Does this include research organizations and universities that have conducted

surveys, but surveys are not the primary research focus? It should if they undertake the surveys in a professional manner.

NOAA requires the survey organization to have implemented procedures to meet the standards outlined in the Council of American Survey Research Organization's (CASRO) Code of Standards for Survey Research, or the American Association for Public Opinion Research's (AAPOR) Code of Professional Ethics and Practices.[17] NOAA should clarify that the CASRO and AAPOR code of ethics are examples of professional conduct code of ethics. For example, Canadian organizations may assist on projects that straddle international borders and may subscribe to Rules of Conduct and Good Practice from the Professional Marketing Research Society of Canada (PMRSC), which are similar.

NOAA should not require that the organization have already implemented practices similar to those mentioned, but that such practices should be implemented for NRDAs. We will be happy to supply NOAA the AAPOR code of professional ethics and practices to consider and incorporate.

2.8. MULTIPLICATIVE NATURE OF COST FACTORS

As McClelland and Schulze discuss in Appendix B, we believe that all of the estimated cost increases shown in Sections 2.3.1, 2.3.2, 2.4, and 2.6 are largely if not entirely multiplicative. In other words, in order to calculate the increased costs implied by all four of the NOAA requirements / guidelines, one should multiply all the cost ratio factors discussed above together. We acknowledge, however, that multiplicativeness may not strictly hold for each and every aspect of each requirement / guideline. It therefore might be reasonable in trying to determine a best available conservative estimate of the overall cost increases from all four regulations / guidelines to use the lower bound (a factor of five) rather than the best available estimate (a factor of 11.67) for the largest contributor to the cost increases, namely, in-person interviews.

[17] Preamble page 1145, column 1 at line 45.

3. INCLUDE FULL AND ACCURATE PASSIVE USE DAMAGES

3.1. WHY IT SHOULD BE EASIER TO INCLUDE FULL PASSIVE USE DAMAGES

The available information indicates that passive use damages represent a very significant proportion of total NRDs (Brown, 1993). Failure to include these damages in NRDAs would result in failing to make the public whole and failing to provide optimum incentives for avoiding NRDs. That, in turn, would make EPA's job more difficult because more pollution would occur than would otherwise be the case or than would be economically efficient. It is therefore very important that the regulations make every possible effort to include passive use losses in NRDs.

3.2. EMPHASIZE ACCURACY AND CREDIBILITY

Accuracy and credibility are very important issues for CV results and CV studies need to incorporate reasonable measures to ensure accuracy and to test for credibility of the results. Several recommendations included by NOAA speak to these concerns and EPA agrees with many of these recommendations as general guidelines, although EPA believes that detailed prescriptions for how these are implemented should be removed. These recommendations include testing for consistency in responses to different questions, respondent comprehension, scope sensitivity, context sensitivity, cognizance of budget constraints and substitution options, and credibility of zero and very high WTP responses. Researchers should be encouraged to consider these issues when designing and analyzing CV studies, but the implementation details for specific accuracy and credibility testing should be left to the specific case and be flexible to incorporate new research findings.

The survey research literature is replete with examples of how the wording of an instrument can influence responses. Downward bias and upward bias can be purposefully introduced in a CV instrument designed to elicit WTP estimates, but neither is desirable. Accuracy, clarity and balance should be the goal of the instrument design, not elicitation of high or low WTP responses.

3.3. ENCOURAGE RATHER THAN DISCOURAGE TRUSTEES TO VALUE PASSIVE USES

The NOAA proposed regulations provide for several different approaches towards estimating natural resource damages. Only one of these, Comprehensive Damage Assessment (CDA), involves estimation of passive use damages. Including such damages in other approaches would result in a proportionate increase in the actions by those using natural resources to avoid or reduce damages.

3.4. MODIFY COMPENSATION FORMULAS TO INCLUDE PASSIVE USE DAMAGES

We recognize that even with added flexibility in the regulations, there will always be smaller NRD cases where it is not economically feasible to measure the damages. In these smaller cases we are concerned that the current compensation formulas do not include passive use losses. Since these are the values that are most likely to be used in the majority of cases, the practical result is that passive use losses are effectively excluded from the damages in most cases. The net effect is to value such damages as zero, which is inappropriate in our view. We suggest that a serious effort be made to include such damages in the compensation formulas.

One approach to doing so involves using a ratio of passive use value to use value based on a review of prior studies. One recent review of empirical studies that have estimated both values (Brown, 1993) found that the ratio of passive use values to use values averaged about 1.9 to 1. In other words, for every dollar of use value lost, 1.9 times that amount of passive use value is lost on average. Although it is likely that this ratio varies substantially with the "commodity" involved, many of the studies reviewed dealt with resources that are adversely affected by oil spills. Specifically, of the 31 studies reviewed by Brown, 11 dealt with fish and wildlife while another ten dealt with water quantity or quality. Another study (Silberman, *et al.*, 1992) dealt with beaches. Although the ratio approach is at best an approximation of the value in each case, so is the rest of the use values determined by the current compensation formulas. Because of the fact that this ratio is likely to change over time as more research is done, and the possibility that a better approach will be found, it is important that whatever is done on this issue be done in such a way that it can be modified reasonably rapidly. One such way might be to include the actual numerical ratio to be used in the proposed guidance document rather than in the regulation. Obviously, this is a high priority for further research since it will have such a major impact on total damages.

3.5. REMOVE THE 50 PERCENT CALIBRATION FACTOR

There are several recommendations in the regulations that "conservative" approaches should be chosen whenever possible in the CV design and interpretation of CV results. This is given in some places as a general recommendation such in the NOAA regulations as "the trustee(s) is encouraged to choose that alternative that would understate the natural resource damages rather than overstate the damages."[1] In other cases, there are specific prescriptions proposed that would result in lower rather than higher damage estimates.

The proposed regulations suggest that a 50 percent, or some other selected percentage, calibration factor be applied to CV results for passive use values. This is apparently based

[1] Page 1146.

on the presumptions that CV results are upward biased and that the amount of upward bias can be reasonably estimated. There is not sufficient support in the available CV literature for the apparent presumption that CV results are upward biased. **Requiring an unsupported reduction in CV results by some amount has no clear empirical basis. EPA recommends that the calibration requirement be dropped from the CV regulations.** The objective should be to reasonably minimize all types of known bias in CV study design and execution.

Comments in the preamble itself are inconsistent with the idea of calibration of CV results. On page 1156 NOAA states, "NOAA has found no empirical evidence to support the contention that CV measures of passive use values are so upwardly biased to be punitive." If there is no such substantive upward bias, then a 50 percent calibration is neither supported nor warranted.

On page 1161, NOAA suggests that trustees can use a calibration factor other than 50 percent if they show that another calibration factor is appropriate. This sounds like reasonable flexibility, but it actually holds trustees to a standard that NOAA itself cannot meet. NOAA has not given any empirical evidence on which the 50 percent calibration factor is based, and acknowledges on page 1157 that there are not methods available to provide external validity for passive use values estimated using CV techniques. If determining an appropriate calibration factor to eliminate suspected upward bias in CV results were a straightforward research question that is readily addressed, NOAA would be able to provide a recommended adjustment factor and give the evidence upon which it is based.

As a practical matter, many treatments of CV data to address concerns related to the long tail of high WTP responses that is typically observed result in significantly lower mean WTP estimates. If calibration remains a required step in analysis of CV results, the effect of data treatments to address concerns about the high-end tail should be counted toward the calibration, which argues against the need for a downward calibration factor when these treatments are used. However, we urge that the regulations recommend some evaluation of the credibility of the responses without stipulating how this is to be done. Sometimes this data treatment is done with evaluation of individual responses. For example, evaluations of WTP responses that are very high relative to the sample means sometimes reveals lack of credibility in some of these responses relative to income levels reported by respondents. An alternative approach is to weight responses on the presumption that the error distribution is log-normal (*i.e.*, higher WTP responses are subject to larger error). Yet another approach is to simply drop a pre-determined number of responses at the highest and lowest ends of the distribution on the presumption that answers at the extremes are less credible than answers closer to the mean.

One of the key concerns in the CV literature that some have used to argue for a downward calibration factor is that WTP responses to hypothetical questions may be

upward biased because respondents are stating what they would pay and are not actually required to pay at that time. This is a particular concern with regard to commodities that may be perceived as good causes, such as protection of endangered species. These are reasonable and legitimate concerns, but the empirical evidence available at this time is inadequate to determine whether significant upward bias exists in responses to hypothetical questions or to determine the expected magnitude of such bias.

Mitchell and Carson (1989) review several studies that compare WTP responses to hypothetical questions to actual payments made in a corresponding simulated market for private or semi-private goods. Studies included in this review were Bohm (1972), Bishop and Heberlein (1986), Bishop *et al.* (1983), Bishop *et al.* (1984), Bishop and Heberlein (1985), Heberlein and Bishop (1986), and Dickie *et al.* (1987). Mitchell and Carson conclude, "For goods which are well understood by respondents (hunting permits, admission to see a TV show), the correspondence between hypothetical and simulated was shown to be quite strong." (page 208) They note, however, that these findings are not directly applicable to the use of CV to value public goods.

Recent experimental studies have compared responses to hypothetical WTP questions, primarily of an open-ended format, to results of actual auctions, primarily Vickrey auctions) designed to reveal maximum WTP. These studies include Coursey *et al.* (1987), Boyce *et al.* (1989; 1992), Irwin *et al.* (1992), McClelland and Schulze (1993), and Neill *et al.* (1994). Again, all of these experiments are for private goods, and the findings are not necessarily applicable to public goods. Two of these studies found the mean hypothetical WTP to be very similar to the mean auction bid in magnitude; two studies found mean hypothetical WTP to be equal to or as much as two times larger than the mean auction bid in different versions of the experiment, and one study found mean hypothetical WTP to be four to nine times the mean auction bid. These results show some potential for apparent upward bias in the hypothetical responses, but the presence and magnitude of the bias varies considerably.

There is simply not enough empirical evidence available at this time to conclude whether and at what magnitude upward bias exists in hypothetical WTP responses. Requiring that some calibration be performed, even if trustees are able to conduct their own study to determine the calibration level (page 1183), is ill-advised. No single experiment to measure hypothetical bias is definitive. Studies to measure potential bias in hypothetical responses for public goods and passive use values are very difficult to design, because actual markets are very hard to simulate for public goods, especially when passive use values are an important factor. Comparisons of CV results to results of techniques based on observed behavior, such as travel cost, are limited to use values and the findings may not apply for passive use values. Even for private goods, the weight of evidence regarding whether there is bias in hypothetical responses is inconclusive at the present time.

3.6. AVOID DICTATING USE OF LUMP-SUM PAYMENTS

NOAA suggests that lump-sum (one-time) payment vehicles would be preferable to annual payment vehicles under the general recommendation to take a conservative approach.[2] NOAA acknowledges that there is no theoretical basis on which to determine whether a lump-sum or payments over time is the more appropriate way to phrase a WTP question and makes the recommendation that lump-sum be used based on the presumption that it would result in lower WTP estimates and is thus a "conservative" choice in survey instrument design. There may be cases for which a lump-sum payment vehicle is appropriate, but this should not be recommended for all cases. In many cases, it is likely that using a lump-sum payment vehicle would be more than conservative; it would be likely to result in significant downward bias in NRDA estimates. EPA therefore recommends that the preference for lump-sum payment vehicles be dropped.

Because environmental commodities often provide a flow of services over time, the benefits to an individual are often experienced over an extended period of time. The assumption implicit in eliciting lump-sum values is that individuals can make net present value calculations of the expected benefit they will derive from the commodity over their entire lifetime (and possibly beyond with existence and bequest values). Forcing respondents to estimate what they would pay as a one-time payment right now to obtain that benefit over an extended period of time could result in an understatement of the stream of benefits for several reasons including:

- Uncertainty about the future could cause the individual to heavily discount potential benefits in future years.
- The tendency of individuals to have somewhat short time horizons in terms of private consumption decisions does not necessarily imply high implicit discount rates for environmental benefits in future years.
- Current income and credit constraints may limit the perceived ability to pay an amount equivalent to the present value of an annual stream of payments.
- Lack of realism in a hypothetical lump-sum payment vehicle for a service flow that would more realistically be paid over time, such as through higher gas prices or higher taxes, could result in more scenario rejection.

NOAA is proposing that the U.S. Treasury rate should be used for discounting a trustee damage claim (page 1074). If individuals respond to lump-sum value questions with implicit rates of time preference that are significantly greater than the U.S. Treasury rate this could result in inconsistent treatment relative to items that are discounted at the U.S. Treasury rate. The rate of time preference is a measure of the rate at which an individual

[2] Page 1146.

will trade off current consumption for future consumption or visa versa. There is ample evidence from NRDA research and other evidence on consumer behavior that individual rates of time preference for private goods are often much higher than typical market interest rates while rates of discount for environmental goods tend to be low.

Rowe *et al.* (1992) report results of a CV study concerning the Nestucca oil spill in which a lump-sum WTP and an annual WTP for five years were asked in different survey versions. A 10 percent discount rate would imply an expected lump-sum payment to be 3.8 times an annual payment. Rowe *et al.* found the mean lump-sum WTP to be 2.8 times the mean annual WTP. This is closer to a 20 percent discount rate. This might occur if there were a 10 percent financial discount rate plus a 10 percent chance of moving out of the area each year. Whatever the reason, it appears that lump-sum payment questions elicit lower WTP estimates than annual WTP questions, even after accounting for a reasonable financial discount rate. These results refute the argument by Kahneman and Knetsch (1992) that respondents may not actually consider the period of payment and will provide the same state WTP regardless of whether they are asked for an annual payment or a lump-sum payment.

Many examples of relatively high rates of time preferences are available in the literature on consumer behavior. In an econometric study, Hausman (1979) found an implicit discount rate of about 20 percent for individuals making tradeoffs between higher capital costs for energy efficiency and higher operating costs on the purchase of a consumer durable (air conditioners). The discount rate was conversely related to income suggesting lower income individuals (assumed also to be credit constrained) have a higher rate of time preference. Hartman and Doane (1986) estimated an econometric model of the decision to undertake energy saving weatherization versus continued higher heating and cooling costs and found discount rates inversely related to income levels with implicit discount rates as high as 87.8 percent for low income groups. Thaler (1981) found discount rates ranging from one percent to 345 percent using a hypothetical survey valuing potential gains and losses. He found that the discount rate varied inversely with the length of time and the size of the potential gain or loss and that the discount rate on losses was much smaller than for gains.

It is also possible that lump-sum payment questions elicit lower WTP estimates because people may respond to lump-sum valuation questions based on current income without accounting for an intertemporal capital market. CV questions eliciting lump-sum bids do not in general emphasize that the payment could be made over multiple periods through appropriate action in credit markets. In essence, individuals may make a lifetime consumption expenditure decision considering only the current period budget constraint, thus significantly understating their lifetime WTP for the commodity. This hypothesis is consistent with the high rates of time preference reported above.

Finally, lump-sum payment vehicles may not be consistent with NOAA's stated goal that “the method of elicitation should be one with which people are familiar and one which

provides a realistic context in which respondents can choose to increase levels of public goods." (page 1144) Payment programs such as periodic utility bills or increased prices for gasoline (consumed nearly continuously) may be more realistic payment vehicles. Asking individuals to make lump-sum payments for a commodity which would in reality be paid for over time could generate scenario rejection.

3.7. ELIMINATE PREFERENCE FOR REFERENDUM FORMAT APPROACHES

As discussed in Section 2.4, the cost of using a referendum format approach as advocated by NOAA is much higher than for many other approaches. Although NOAA asserts several advantages to the referendum format, there is no empirical evidence given that the results are measurably more accurate than those obtained using other CV question formats. In fact, there is some evidence that the referendum approach may result in an upward bias in the results.

NOAA argues that the yes-no referendum format is advantageous because it is similar to most consumer decisions in which individuals decide whether or not to buy at posted prices. It is clear that a yes-no question format is probably easier for respondents to answer than formats in which they must choose a dollar amount from a payment card or give a response to an open-ended question, but it is not clear that the information obtained is significantly better. Evidence from laboratory economics experiments shows that, in early rounds, posted-offer markets (which have a yes-no format) produce WTP values above the equilibrium price (Davis and Holt, 1993). In summarizing the extensive literature on posted-offer markets, Davis and Holt state, "Price convergence from above and low early period efficiencies are predominant features of posted-offer markets" and "efficiencies...are low, at least relative to efficiencies for most double auctions" (page 179). In summary, there is some evidence that single yes-no referendum questions may elicit WTP estimates that are too high.

NOAA asserts that the referendum format is *incentive compatible*, in that there is no incentive to the respondent to answer inaccurately. This statement is not technically correct. It is generally accepted in the public choice literature that individuals voting in a majority rule referendum will vote yes if it is in their own interest to do so and no if it is not. This implies that the referendum format is theoretically *demand revealing*. Thus, from a theoretical perspective, individuals answering a referendum question are given no incentive to strategically bias their responses. This property, that the mechanism be *demand revealing*, is the one that is of interest in obtaining values, not *incentive compatibility*. Some CV studies have used incentive compatible public good mechanisms such as the Groves-Ledyard or Smith auction (for a discussion of incentive compatible public good mechanisms see Davil and Hold, 1993). However, a majority rule referendum is not incentive compatible since it can lead to an inefficient provision of public goods. There is actually very little evidence that any CV question formats elicit significant strategic bias. Mitchell and Carson (1989) conclude that intentional strategic bias is not likely to be a significant problem in most CV

studies. A more important concern is whether information provided in the question leads the respondents to answer in some way inconsistent with their true WTP. For example, there is evidence that maximum WTP estimates elicited using an iterative bidding format are influenced by the dollar amount offered in the first question (Rowe and Chestnut, 1983).

There is evidence that referendum formats also may not be entirely neutral. For example, Cooper and Loomis (1992) and Cameron and Huppert (1991) found that mean WTP estimates based on referendum questions were sensitive to the ranges and intervals of dollar amounts included in the CV questions. Kanninen and Kristrom (1993) show that sensitivity of mean WTP to bid values can be caused by model misspecification, failure to include bid values that cover the middle of the distribution, or inclusion of bids from the tails of the distribution. These findings suggest that results based on the referendum format may not be robust, particularly when procedures to determine the appropriate range and allocation of dollar amounts are not well-established.

These findings of different mean WTP estimates when different ranges and intervals are used in the referendum approach suggest that respondents may be influenced by the dollar amount offered in the referendum question. One possible difficulty is that respondents who would like to see the good provided, but who find the dollar amount too high, may feel some incentive to say yes anyway in order to register their desire to have the good provided. Their only opportunity to affirm their support for the good in question is to say yes to the offered amount, even if it is higher than they actually would be willing to pay. Results reported by Bishop and Boyle (1985) for a referendum format regarding whether to establish a nature preserve in an Illinois state park support this hypothesis. A significant share of respondents indicated in response to follow-up questions that they really did not know how much they would be willing to pay for the nature preserve, but that they gave a yes vote for the amount in the question because they thought establishing the preserve was a good idea.

Several studies have made direct comparisons of mean WTP estimates obtained using a referendum format versus an open-ended format. These studies and their findings are summarized in Table 3-1. These results suggest that mean WTP estimates based on referendum questions are similar to or higher than estimates based on open-ended questions. It remains uncertain whether there is a tendency for upward bias in the referendum format or downward bias in the open-ended format, because the “true” value is unknown. A detailed review of these studies would reveal limitations in each of the comparisons, but as a whole, these studies do not demonstrate that the referendum format is necessarily superior to other WTP question formats that might be used. Combined with the experimental and other evidence cited above, there is reason to suspect that there may be some upward bias in the referendum results. The results of these comparisons do suggest that there is unlikely to be more upward bias in open-ended format results than in referendum results.

3.8. AVOID USING THRESHOLD OR SCREENING FACTORS

We fail to understand the utility of using screening factors to determine whether trustees should carry out a CV study. Presumably the trustees, especially given their knowledge of their own budgetary situation, are best equipped to decide on their own whether or not to carry out a CV study. Further, they have ample incentives to exercise good judgment. They will want to at least recover the cost of such studies, and are unlikely to undertake such studies if they do not believe that they can do so. In addition, trustees must be able to withstand judicial review of their costs. Finally, proposed screening factors are difficult to support unless they are based on careful research concerning their effect on trustee behavior, which would be difficult to conduct. The most likely result of requiring such screening factors will be to further decrease the possibilities for using CV and therefore for including passive use values in NRDs.

Table 3-1: COMPARISON OF THE DICHOTOMOUS CHOICE AND OPEN-ENDED QUESTION FORMATS IN THE CONTINGENT VALUATION METHOD

Author(s)	Commodity	Ratio of Mean Dichotomous-Choice Values to Open-Ended
Johnson, R.L., N.S. Bregenzer, and B. Shelby. 1990.	Non-commercial river rafting on Oregon's Rogue river.	1.62
Jordan, J.L. and A.H. Elnagheeb. 1994.	Synthetic Monte Carlo experiment—no commodity Compared payment-card to DC	Depended on sample size (Table 8) n=100 1.62 n=200 1.31 n=400 1.40 n=600 1.33
Kriström, B. 1993.	Preserving forest areas	Means not reported[3]
Walsh, R.G., D.M. Johnson, and J.R. McKean. 1989. _____. 1992.	Outdoor recreation	1.3
Kealy, M.J., J.F. Dovidio, and M.L. Rockel. 1988.	Public good—water ecosystem quality in Adirondacks	1.4-2.5
Kealy, M.J. and R.W. Turner. 1993	Private good—candy bar	1.0
Seller, C., J.R. Stoll, and J. Chavas. 1985	Recreational boating permits	4.8 to 9.5 across the various lakes
Boyle, K.J. and R.C. Bishop. 1988.	Scenic beauty for Wisconsin boaters and canoers Comparison of iterative bidding, payment card, and dichotomous choice	3.0 (prior to truncation experiment p. 25)

[3] Kriström (1993) used a non-parametric test to compare the distributions of the results from the different formats and could not reject the null hypothesis that the values were the same.

4. AVOID PREMATURELY FREEZING CV METHODOLOGY AND GENERAL OBSERVATIONS

4.1. WHY IT IS IMPORTANT TO AVOID FREEZING CV METHODOLOGY

Attempts to codify in regulation what constitutes "good" methodology run the risk that they will result in freezing the methodology that is used at the point that the regulations are written, even assuming that they represent the best methods available at that time. They may also inhibit the search for new and improved methodologies since there would be no rebuttable presumption for assessments using them for NRDA purposes. It is important that trustees be encouraged to adopt methodological improvements so that their studies will represent the current state-of-the-art in reliability, rather than continuing to use methodologies that may become outmoded.

The evolution of the travel cost method for valuing recreation resources over the last 10 to 15 years illustrates the negative effect that might have occurred if prescriptive regulations had been established based on the best methods being used 10 to 15 years ago. When the U.S. Water Resource Council (1982) guidelines on valuation of recreational resources were drafted, the travel cost methods being implemented were zonal models best able to estimate consumer surplus values for the existence of a recreational site, such as a reservoir. The models were typically not able to determine values for differences in the characteristics of the site, such as water quality or fishing catch rates, and did not distinguish between average population characteristics in the origin zone. These models were considered best practice at the time and if their application had been dictated by strict prescriptive regulations many important innovations in the travel cost method would probably have been hindered. Following some early innovations suggested and implemented by Bockstael *et al.* (1987), Greig (1983), Brown *et al.* (1983), Morey (1981), and others, travel cost methods now routinely use individual specific data to estimate models consistent with traditional consumer demand theory. Important analytical innovations include incorporation of multiple recreation sites, quality attributes, substitute activities, individual preferences and characteristics, and the participation decision. Applications of the travel cost method now produce far more accurate and appropriate value estimates for many policy issues than the unit value estimates (average value per recreation day) that were the focus of the Water Resource Council review based on best available analysis at that time.

4.2. GENERAL OBSERVATIONS

One approach to solving these problems would be to make all the changes proposed earlier in these comments without changing the basic structure of the regulations. Specifically, this would do the following:

- Reduce the cost of contingent valuation studies
  - Use cost-effective approach to selecting CV requirements

- Use within rather than between subject performance tests
- Eliminate the scope test as now defined
- Eliminate preference for referendum format approaches
- Require more modest response rates
- Use lower cost survey techniques
- Drop requirement for use of survey research organization

- Include full and accurate passive use damages
  - Emphasize accuracy and credibility
  - Encourage rather than discourage trustees to value passive uses
  - Modify compensation formulas to include passive use damages
  - Remove the 50 percent calibration factor
  - Avoid dictating use of lump-sum payments
  - Eliminate preference for referendum format approaches
  - Avoid using threshold or screening factors.

Such changes, would, in EPA's view, greatly improve the regulations, but would leave the substantial possibility that the regulations would freeze the state of the CV art in unforeseen ways. It is difficult if not impossible to foresee every scientific advance that will be made. Therefore, EPA advocates as flexible regulations as possible, avoiding unnecessarily prescriptive standards, and putting as many of the requirements as possible in the proposed guidance document. In this regard, we strongly believe that much of the restrictive language concerning design standards which is contained in the Preamble should be removed. Failing that, we suggest that the Preamble be at least caveated so as to make it clear that NOAA/DOI recognize there are other approaches for doing CV that may produce equally accurate results.

REFERENCES

Ayidiya, S.A. and M.J. McCiendon. 1990. "Response Effects in Mail Surveys." *Public Opinion Quarterly* 54: 229-247.

Babbie, E. R. 1973. *Survey Research Methods*. Wadsworth, Belmont, CA.

Bailly, K. D. 1978. *Methods of Social Research*. Free Press, New York

Baird, J. C., and E. Noma. 1979. *Fundamentals of Scaling and Psychophysics.* Wiley, New York, pp. 25-47.

Bishop, G. G., H. Hippler and F. Schwarz. 1988. "A Comparison of Response Effects In Self Administered and Telephone Surveys." In Groves *et al.*, eds., *Telephone Survey Methodology*. New York: John Wiley and Sons.

Bishop, R.C., and K.J. Boyle. 1985. *The Economic Value of Illinois Beach State Nature Preserve*. Final report prepared for the Illinois Department of Conservation.

Bishop, R.C., and T.A. Heberlein. 1985. "Progress Report of the 1984 Sandhill Study." Preliminary report to the Wisconsin Department of Natural Resources.

Bishop, R.C., and T.A. Heberlein. 1986. "Does Contingent Valuation Work?" in R.G. Cummings, D.S. Brookshire, and W.D. Schulze (eds.), *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa NJ: Rowman & Allanheld.

Bishop, R.C., T.A. Heberlein, and M.J. Kealy. 1983. "Contingent Valuation of Environmental Assets: Comparisons with a Simulated Market," *Natural Resources Journal* 23(July):619-34.

Bishop, R.C., T.A. Heberlein, M.P. Welsh, and R.A. Baumgartner. 1984. "Does Contingent Valuation Work? A Report on the Sandhill Study." Paper presented at the Joint Meeting of the Association of Environmental and Resource Economists and the American Economics Association, Cornell University, Ithaca NY, August.

Bockstael, N.E., K.E. McConnell, and I.E. Strand. 1987. *Benefits from Improvements in Chesapeake Bay Water Quality*—Vol. II, prepared by Department of Agricultural and Resource Economics, University of Maryland for US Environmental Protection Agency, Washington, DC.

Bohm, P. 1972. "Estimating Demand for Public Goods: An Experiment." *European Economic Review* 3:111-130.

Boyce, R., G.H. McClelland, T. Brown, G. Peterson, and W.D. Schulze. 1989. *Economic*

Explanation of the Empirical Disparity Between Willingness to Pay (WTP) and Willingness to Accept (WTA) Compensation: Do Existence Values Exist? Final report submitted to the U.S. Forest Service by the University of Colorado.

Boyce, R., G.H. McClelland, T. Brown, G. Peterson, and W.D. Schulze. 1992. "An Experimental Examination of Intrinsic Values as a source of the WTA-WTP Disparity." *American Economic Review* 82(5):1366-1373.

Boyle, K.J. 1990. "Dichotomous-Choice, Contingent-Valuation Questions: Functional Form is Important," *Northeastern Journal of Agricultural and Resource Economics,* October: 125-131.

Boyle, K.J., and R.C. Bishop. 1988. "Welfare Measurements Using Contingent Valuation: A Comparison of Techniques," *American Journal of Agricultural Economics* 70(1):20-28.

Bregenzer, and B. Shelby. 1990. "Contingent Valuation Question Formats: Dichotomous Choice versus Open-Ended Responses," in *Economic Valuation of Natural Resources: Issues, Theory, and Applications*, eds. R.L. Johnson and G.V. Johnson, Westview Press, Boulder, CO.

Brown, Thomas. 1993. "Measuring Nonuse Values: A comparison of Recent Contingent Valuation Studies." In John Bergstrom, Benefits and Costs of Transfer in Natural Resources Planning, Sixth Interim Report W-133, Department of Agricultural Economics, University of Georgia, Athens.

Brown, W.G., C. Sorkus, B. Chou-Yong, and S. Richards. 1983. "Using Individual Observations to Estimate Recreational Demand Functions," *American Journal of Agricultural Economics*, Vol 65, February, pp. 154-57.

Cameron, T.A., and D.D. Huppert. 1989. "OLS versus ML Estimation of Non-market Resource Values with Payment Card Interval Data," *Journal of Environmental Economics and Management* 17:230-246.

Cameron, T.A., and D.D. Huppert. 1991. "Referendum Contingent Valuation Estimates: Sensitivity to the Assignment of Offered Values," *Journal of American Statistics Association*, 86(416):910-918.

Cameron, T.A., and M.D. James. 1987. "Efficient Estimation Methods for 'Closed-Ended' Contingent Valuation Surveys." *Review of Economics and Statistics* 69(2):269-276.

Chestnut, Lauraine G. and Robert D. Rowe. 1990. "Review and Response to: 'Development and Design of a Contingent Value Survey For Measuring the Public's Value for Visibility Improvements at the Grand Canyon National Park' September, 1990 Revised Draft Report by Decision Focus Incorporated." RCG / Hagler Bailly report to the Economic Analysis

Branch, Office of Air Quality Planning and Standards, U.S. EPA. Research Triangle Park, North Carolina. December 10. (Also in the Navajo Generating Station Case EPA administrative Record)

Cooper, J. and J. Loomis. 1992. "Sensitivity of Willingness-to-Pay Estimates to Bid Design in Dichotomous Choice Contingent Valuation Models," *Land Economics* 68(2):211-224.

Cooper, J. and J. Loomis. 1992. "Sensitivity of Willingness-to-Pay Estimates to Bid Design in Dichotomous Choice Contingent Valuation Models," *Land Economics* 68(2):211-224.

Coursey, D.L., J.L. Hovis, and W.D. Schulze. 1987. "The Disparity Between Willingness to Accept and Willingness to Pay Measures of Value." *Quarterly Journal of Economics* 679-685.

Cummings, R.G., P.T. Ganderton, and T. McGucking. 1994. "Substitution Effects in CVM Values," *American Journal of Agricultural Economics* 76:205-214.

Davis, D.D., and C.A. Holt. 1993. *Experimental Economics.* Princeton, NJ: Princeton University Press.

Dickie, M., A. Fisher, and S. Gerking. 1987. "Market Transactions and Hypothetical Demand Data: A Comparative Study." *Journal of Environmental Economics and Management* 5(1):63-80.

Dillman, D. A. 1978. *Mail and Telephone Surveys: The Total Design Method.* New York: John Wiley & Sons, New York.

Dillman, D. A. 1983. "Mail and Other Self-administered Questionnaires." in P. Rossi, J. Wright and A. Anderson (eds.), *Handbook of Survey Research.* New York, Academic Press.

Dillman, D. A., Jon R. Clark, and James B. Treat. 1994. "Influence of 13 Design Factor on Completion Rates to Decennial Census Questionnaires." Paper presented at the 1994 Annual Research Conference of the U.S. Bureau of the Census, Key Bridge Marriott Hotel, Arlington, VA, March 21.

Doyle, James K., Gary H. McClelland, William D. Schulze, Steven R. Elliott, and Glenn W. Russell. 1991. "Protective Responses to Household Risk: A Case Study of Radon Mitigation," *Risk Analysis*, Vol. 11, No. 1, pp. 121-34.

Duffield, J.W., and D.A. Patterson. 1991. "Inference and Optimal Design for a Welfare Measure in Dichotomous Choice Contingent Valuation," *Land Economics* 67(2):225-39.

Erlebacher, A. 1977. "Design and Analysis of Experiments Contrasting the within- and between-Subjects Manipulations of the Independent Variable," *Psychological Bulletin*, Vol. 84, pp. 212-9.

Fisher, Ann, and Robert Raucher. 1984. "Intrinsic Benefits of Improved Water Quality: Conceptual and Empirical Perspectives." In V. K. Smith and A. Dryden, eds., *Advances in Applied Microeconomics*, Vol. 3. JAI Press.

Goyder, J. 1985. "Face-to-Face Interviews and Mailed Questionnaires: The Net Difference in Response Rate." *Public Opinion Quarterly 49:234-252.*

Grieg, P.J. 1983. "Recreation Evaluation using a Characteristics Theory of Consumer Behavior," *American Journal of Agricultural Economics*, Vol. 65, February, pp. 90-97.

Hartman, R.S., and M.J. Doane. 1986. "Household Discount Rates Revisited." *The Energy Journal* 7:139-148.

Hausman, J.A. 1979. "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables." *The Bell Journal of Economics* 10(Spring):33-54.

Heberlein, R.A., and R.C. Bishop. 1986. "Assessing the Validity of Contingent Valuation: Three Field Experiments." *Science of the Total Environment* 56:99-107.

Hoehn, J.P., and A. Randall. 1987. "A Satisfactory Benefit Cost Indicator from Contingent Valuation," *Journal of Environmental Economics and Management* 14:226-247.

Irwin, J.R., G.H. McClelland, and W.D. Schulze. 1992. "Hypothetical and Real Consequences in Experimental Auctions for Insurance Against Low-Probability Risks." *Journal of Behavioral Decision Making* 5:107-116.

Jordan, J.L., and A.H. Elnagheeb. 1994. "Consequences of Using Different Question Formats in Contingent Valuation: A Monte Carlo Study," *Land Economics* 70(1):97-110.

Johnson, R.L., N.S. Bregenzer, and B. Shelby. 1990. "Contingent Valuation Question Formats: Dichotomous Choice versus Open-Ended Responses," in *Economic Valuation of Natural Resources: Issues, Theory, and Applications*, eds. R.L. Johnson and G.V. Johnson, Westview Press, Boulder, CO.

Kahneman, D., and J.L. Knetsch. 1992. "Valuing Public Goods: The Purchase of Moral Satisfaction." *Journal of Environmental Economics and Management* 22(1):57-

Kanninen, Barbara J. 1993. "Optimal Experimental Design for Double-bounded Dichotomous Choice Contingent Valuation," Land Economics, May, Vol. 69, pp. 138-46.

Kanninen, Barbara J. and Bengt Kristrom. 1993. "Sensitivity of Willingness-to-Pay Estimates to Bid Design in Dichotomous Choice Valuation Models: Comment," *Land Economics*, Vol. 69, No. 2, May, pp.199-202.

Keren, G. 1993. "Between- or within-Subjects Design: A Methodological Dilemma." In G. Keren and C. Lewis, eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Erlbaum, Hillsdale, NJ.

Kealy, M.J., J.F. Dovidio, and M.L. Rockel. 1988. "Accuracy in Valuation is a Matter of Degree," *Land Economics*, 64(2):158-171.

Kealy, M.J. and R.W. Turner. 1993. "A Test of the Equality of Closed-Ended and Open-Ended Contingent Valuations," *American Journal of Agricultural Economics* 75(May):321-331.

Kriström, B. 1993. "Comparing Continuous and Discrete Contingent Valuation Questions," *Environmental and Resource Economics* 3:63-71.

Krysan, Maria, Howard Schuman, Lesli Jo Scott, and Paul Beatty. Forthcoming. "Mail versus Face to Face Survey: A Comparison of Response Rates and Response Content Based on a Probability Sample," *Public Opinion Quarterly*.

Lepkowski, J.M. 1987. "Telephone Sampling Methods the United States." in M.L. Groves et al. (eds.), *Telephone Survey Methodology*. New York, Wiley.

Loomis, John. 1987. "Expanding Contingent Value Sample Estimates to Aggregate benefit Estimates: Current Practices and Proposed Solutions." *Land Economics* 63 (4):396-402.

Loomis, J.B. 1990. "Comparative Reliability of the Dichotomous Choice and Open-Ended Contingent Valuation Techniques," *Journal of Environmental Economics and Management* 18:78-85.

McClelland, G.H., and W.D. Schulze. 1993. "Insurance for Low-Probability Hazards: A Bimodal Response to Unlikely Events." *Journal of Risk and Uncertainty* 7:95-116.

McClelland, Gary, William Schulze, and Edward Balistreri. 1994. *An Examination of Performance Testing Requirements for Contingent Valuation*. Draft report to U.S. EPA by the University of Colorado. Included as Appendix B of these comments.

McClelland, Gary, William Schulze, Donald Waldman, and D. Jay Goodman. 1994a. *An Examination of the Proposed Scope Test using Market Data*. Draft report to U.S. EPA by the University of Colorado. Included as Appendix C of these comments.

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

Morey, E. 1981. "The Demand for Site-Specific Recreational Activities: A Characteristics Approach," *Journal of Environmental Economics and Management*, Vol. 8, pp. 345-71.

Neill, H.R., R.G. Cummings, P.T. Ganderton, G.W. Harrison, and T. McGuckin. 1994. "Hypothetical Surveys and Real Economic Commitments," *Land Economics* 70(2):145-54.

Rosenthal, R., and D. Rubin. 1980. "Comparing within- and between-Subjects Studies," Sociological Methods and Research, Vol. 9, pp. 127-36.

Rowe, R.D., and L.G. Chestnut. 1983. "Valuing Environmental Commodities: Revisited." *Land Economics* 59: 404-410.

Rowe, R.D., W.D. Shaw and W. Schulze. 1992. "Nestucca Oil Spill." K.M. Ward and J. W. Duffield (eds.), *Natural Resource Damages: Law and Economics.* New York, John Wiley & Sons, Inc. pp. 527-554.

Schulze, William D., Robert D. Rowe, William S. Breffle, Rebecca Boyce and Gary McClelland. 1993. *Contingent Valuation of Natural Resource Damages Due to Injuries to the Upper Clark Fork River Basin.* RCG/Hagler Bailly report to the State of Montana Natural Resource Damage Program.

Seller, C., J.R. Stoll, and J. Chavas. 1985. "Validation of Empirical Measures of Welfare Change: A Comparison of Nonmarket Techniques," *Land Economics* 61(2):156-175.

Shuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Format, Wording and Content.* New York, Academic Press.

Silberman, Jonathan, Daniel Gerlowski, and Nancy Williams. 1992. "Estimating Existence Value for User and Non-users of New Jersey Beaches," *Land Economics*, 53(2), pp. 185-198.

Smith, V.L. 1979. "Incentive Compatible Experimental Processes for the Provision of Public Goods." In V.L. Smith (ed.) *Research in Experimental Economics*, Vol 1. Greenwich CN: JAI Press.

Smith, V.L. 1980. "Experiments with a Decentralized Mechanism for Public Good Decisions." *American Economic Review* 70:584-599.

Thaler, R. 1981. "Some Empirical Evidence on Dynamic Inconsistency." *Economics Letters* 8:201-207.

Thomberry, 0. and J. Massey. 1987. "Trends in the United States Telephone Coverage Across Time and Subgroups." In M. L. Groves *et al.* (eds.), *Telephone Survey Methodology*. Wiley, New York, NY.

U.S. Environmental Protection Agency. 1991. "Approval and Promulgation of Implementation Plans: Revision of the Visibility FIP for Arizona: Final Rule," *Federal Register*, Vol. 56, No. 192, October 3, p. 50172.

U.S. Environmental Protection Agency, Office of Water. March 1994. *President Clinton's Clean Water Initiative: Analysis of Benefits and Costs*, Report No. EPA-800-R-94-002.

U.S. Water Resources Council. 1983. *Principles and Guidelines for Water and Related Land Resources Implementation Studies.* Washington, D.C.

Walsh, R.G., D.M. Johnson, and J.R. McKean. 1989. "Issues in Nonmarket Valuation and Policy Application: A Retrospective Glance," *Western Journal of Agricultural Economics* 14(1):178-188.

Walsh, R.G., D.M. Johnson, and J.R. McKean. 1992. "Benefit Transfer of Outdoor Recreation Demand Studies, 1968-1988," *Water Resources Research*, 28(3):707-713.

APPENDIX A. PARTIAL BIBLIOGRAPHY OF EPA-SPONSORED CONTINGENT VALUATION STUDIES

This Appendix contains a partial bibliography of EPA-sponsored studies on contingent valuation. It contains studies that have been published or funded by EPA and journal articles that can reasonably be said to have been based on EPA-funded studies. Inclusion on this list does not imply that EPA necessarily endorses the findings reported in these studies.

Ashford, Nicholas A., and Christopher T. Hill. 1982. *Analyzing the Benefits of Health, Safety, and Environmental Regulations,* prepared by Massachusetts Institute of Technology for US Environmental Protection Agency.

Berger, Mark C., Glenn C. Bloomquist, D. Kenkel, and George S. Tolley. 1987. "Valuing Changes in Health Risks: A Comparison of Alternative Measures," *Southern Economic Journal*, Vol. 53, No. 4, pp. 967-984.

Brookshire, David S., D. L. Coursey, and William D. Schulze. 1987. "The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior," *Economic Inquiry*, Vol. XXV, No. 2.

Brookshire, D. S., and R. d'Arge. 1979. "Resource Impacted Communities: Economics, Planning and Management," paper prepared for the 4th U.S.-U.S.S. Symposium on Comprehensive Analysis of the Environment, Jackson, Wyoming, October 21-27, pp. 1-55.

Brookshire, D. S., R. D'Arge, W. Schulze, and M. A. Thayer. 1979. *Methods Development for Assessing Trade-offs in Environmental Management,* US Environmental Protection Agency, EPA, Washington, DC.

Brookshire, David S., Ralph C. D'Arge, William D. Schulze, and Mark A. Thayer. 1979. Methods Development for Assessing Tradeoffs in Environmental Management, Vol. II, "Experiments in Valuing Non-Market Goods: A Case Study of Alternative Benefit Measures of Air Pollution Control in the South Coast Air Basin of Southern California." Report EPA-600/5-79-001b, US Environmental Protection Agency, Washington, DC

Brookshire, D. S., Ralph C. d'Arge, William Schulze, and Mark Thayer. 1982. Experiments in Valuing Public Goods. In *Advances in Applied Microeconomics*, ed. Smith, V. K., JAI Press, Inc., Greenwich, CT, pp. 123-172.

Brookshire, D. S., Ronald G. Cummings, Morteza Rahmatian, and William D. Schulze, *et al.* 1982. *Experimental Approaches for Valuing Environmental Commodities*.

Brookshire, D. S., and T. Crocker. 1981. "The Advantages of Contingent Valuation Methods for Benefit Cost Analysis," *Public Choice*, Vol. 36, No. 2, pp. 235-252.

Brookshire, David S., B.C. Ives, and William D. Schulze, "The Valuation of Aesthetic Preferences," *Journal of Environmental Economics and Management*, Volume 3, pp. 325-46.

Brookshire, D. S., R.C. d'Arge, and W.D. Schulze. 1981. "Valuing Environmental Commodities: Some Recent Experiences." *Land Economics*, May.

Brookshire, D. S., W.D. Schulze. 1983. "The Economic Benefits of Preserving Visibility in the National Parklands of the Southwest," *Natural Resources Journal*, Vol. 23, January.

Brookshire, D. S., W.D. Schulze, M.A. Thayer, and R.C. d'Arge. 1982. "Valuing Public Goods: A Comparison of Survey and Hedonic Approaches," *American Economic Review*, Vol. 72, No. 1, pp. 165-77.

Cameron, Trudy A. 1989. *Contingent Valuation Assessment of The Economic Damages of Pollution to Marine Recreational Fishing*, prepared by University of California at Los Angeles for US Environmental Protection Agency.

Cantor, Robin. 1993. *Community Preferences and Superfund Responsibilities*. Report to the U.S. Environmental Protection Agency, Washington, D.C.

Cameron, Trudy Ann. 1989. *The Effects of Variations in Gamefish Abundance on Texas Recreational Fishing Demand: Welfare Estimates*, prepared by University of California at Los Angeles for US Environmental Protection Agency.

Cameron, Trudy A. 1992. "Combining Contingent Valuation and Travel Cost Data for the Valuation of Nonmarket Goods," *Land Economics*, Vol. 68, No. 3, pp. 302-317.

Carson, Richard T. 1991. *Comments on the Benefit Analysis in the U.S. Environmental Protection Agency's Proposed Navajo Generating Station BART Action*, US EPA.

Carson, Richard T., Mark Machina, and John Horowitz. 1987. *Discounting Mortality Risks*, final technical report to USEPA by University of California at San Diego, La Jolla, CA.

Carson, Richard T., and Robert Cameron Mitchell. 1991. *The Value of Clean Water: The Public's Willingness to Pay for Boatable, Swimmable and Fishable Water*, prepared by University of California at San Diego for US EPA.

Carson, Richard T., and Robert Cameron Mitchell. 1991. "The Value of Clean Water: The Public's Willingness to Pay for Boatable, Fishable, and Swimmable Water," *Water Resources Research*, Vol. 29, No. 7, pp. 2445-54.

Chestnut, Lauraine G., and Robert D. Rowe. 1990. *Preservation Values for Visibility Protection at the National Parks*, prepared by RCG/Hagler, Bailly for US EPA—OAQPS.

Chestnut, Lauraine G., and Daniel M. Violette. 1990. *The Relevance of Willingness-to-Pay Estimates of the Value of a Statistical Life in Determining Wrongful Death Awards*, prepared for EPA Economic Analysis and Research Branch of the US Environmental Protection Agency.

Chestnut, Lauraine G., S. Colome, L. R. Keller, and W. Lambert, *et al*. 1988. *Heart Disease Patients' Averting Behavior, Costs of Illness, and Willingness to Pay to Avoid Angina Episodes —Final Report*, prepared by University of California at Irvine and RCG / Hagler, Bailly, Inc. for US Environmental Protection Agency (OPA), US Government Printing Office, Washington, DC.

Chestnut, Lauraine G., and Daniel M. Violette. 1986. *Estimates of Willingness to Pay for Pollution-Induced Changes in Morbidity: A Critique for Benefit-Cost Analysis of Pollution Regulation*. Environmental Benefits Analysis Series appendix, prepared for EPA Economic Analysis and Research Branch of US Environmental Protection Agency.

Coulson, Amy, Mark Dickie, Shelby Gerking, and William Schulze. 1985. *Improving Accuracy and Reducing Costs of Environmental Benefit Assessments, Volume III: Estimating the Benefits of Reducing Community Low-Level Ozone Exposure.* Draft Report prepared by Center for Economic Analysis of the University of Colorado for US Environmental Protection Agency, Washington, DC.

Coursey, Donald L., David S. Brookshire, Shelby Gerking, Donald Anderson, Mark Dickie, and William D. Schulze. 1986. *Experimental Methods for Assessing Environmental Benefits: Volume II, Laboratory Experimental Economics as a Tool for Measuring Public Policy Values*, report prepared by the University of Wyoming for USEPA.

Coursey, Don L., John Hovis, and William D. Schulze. 1987. "The Disparity between Willingness to Accept and Willingness to Pay Measures of Value," *Quarterly Journal of Economics*, Vol 102, pp. 679-90. Reprinted in John D. Hey and Graham Loomis, ed., *Recent Developments in Experimental Economics*, Edward Elgar Publishing Ltd., 1992.

Coursey, Donald L., and William D. Schulze. 1986. "The Application of Laboratory Experimental Economics to the Contingent Economics to the Contingent Valuation of Public Goods," *Public Choice*, Vol. 49, No. 1, pp. 47-68.

Crocker, Thomas D., John Tschirhart, Richard M. Adams, and Richard W. Katz. Undated. *Methods Development in Measuring Benefits of Environmental Improvements: Volume IV, Valuing Ecosystem Functions: The Effects of Air Pollution*, report prepared by the University of Wyoming for the USEPA.

Cropper, Maureen L., William R. Porter, Berton J. Hansen, Robert A. Jones, and John G. Riley. 1979. *Methods Development for Assessing Air Pollution Benefits, Volume IV, Studies on Partial Equilibrium Approaches to Valuation of Environmental Amenities*, Report EPA-600 / 5-

001d, US Environmental Protection Agency, Washington, DC.

Cummings, Ronald G., David S. Brookshire, and William Schulze. 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*, Rowan and Allanheld, Totowa, NJ.

Cummings, Ronald G., William D. Schulze, Shelby Gerking, and David S. Brookshire. "Measuring the Elasticity of Substitution of Wages for Municipal Infrastructure: A Comparison of the Survey and Wage Hedonic Approaches," *Journal of Environmental Economics and Management*, Vol. 13, No. 3, pp. 269-76.

d'Arge, Ralph C. 1985. *Environmental Quality Benefits Research for the Next Five Years: Some Observations and Recommendations*. 1985. Report to the U.S. EPA from the University of Wyoming.

d'Arge, Ralph C., and Jason Shogren. 1985. *Water Quality Benefits: An Experimental Analysis of the Lakes at Okoboji, Iowa*, prepared by University of Wyoming for US Environmental Protection Agency.

d'Arge, Ralph C., and Jason F. Shogren. 1989. "Non-Market Asset Prices: A Comparison of Three Valuation Approaches," in H. Folmer and E. van Ireland, ed., *Valuation Methods and Policy Making in Environmental Economics*. Elsevier Science Publishers, Amsterdam.

d'Arge, Ralph C., and Jason F. Shogren. 1989. "Okoboji Experiment: Comparing Non-Market Valuation Techniques in Unusually Well-Defined Market for Water Quality," *Ecological Economics*, Vol. 1, No. 3, pp. 251-59.

Desvousges, William H., V. Kerry Smith, Brown, and D. Kirk Pate. 1984. *The Role of Focus Groups in Designing a Contingent Valuation Survey to Measure the Benefits of Hazardous Waste Management Regulations*, prepared by Research Triangle Institute for US EPA - EARB.

Desvousges, William H., V. Kerry Smith, and Matthew P. McGivney. 1983. *A Comparison of Alternative Approaches for Estimating Recreation and Related Benefits of Water Quality Improvements*, Report EPA-230-05-83-001, US Environmental Protection Agency, Washington, D.C.

Desvousges, William, V. Kerry Smith, and Ann Fisher. 1985. *Option Price Estimates for Water Quality Improvements: A Contingent Valuation Study for the Monongahela River*, prepared by Research Triangle Institute, Research Triangle Park, NC for US Environmental Protection Agency.

Desvousges, William, V. Kerry Smith, and Ann Fisher. 1987. "Option Price Estimates for Water Quality Improvements: A Contingent Valuation Study for the Monongahela River,"

Journal of Environmental Economics and Management, Vol. 14, No. 3, pp. 248-267.

Desvousges, William H., and V. Kerry Smith. 1983. An Overview: The Benefits of Hazardous Waste Management Regulations, prepared by RTI for US EPA.

Dickie, Mark, and Shelby Gerking. 1989. "Benefits of Reduced Morbidity from Air Pollution: A Survey of Valuation Methods and Policy Making," in H. Folmer and E. van Ireland, ed., *Valuation Methods and Policy Making in Environmental Economics*, North-Holland, Amsterdam.

Dickie, Mark, Shelby Gerking, William Schulze, Anne Coulson, and Donald Tashkin. 1986. *Value of Symptoms of Ozone Exposure: An Application of the Averting Behavior Method.* Report by the University of Wyoming to the U.S. EPA.

Dickie, M., S. Gerking, D. Brookshire, and D. Coursey, *et al*. 1987. *Reconciling Averting Behavior and Contingent Valuation Benefit Estimates of Reducing Symptoms of Ozone Exposure*. Prepared for US Environmental Protection Agency by University of Wyoming.

Dickie, Mark, Ann Fisher, and Shelby Gerking. 1987. "Market Transactions and Hypothetical Demand Data: A Comparative Study," *Journal of the American Statistical Association*, Vol. 82, No. 397, pp. 69-75.

Dickie, Mark, Shelby Gerking, and Mark Agee. 1990. *Stratospheric Ozone Depletion, Skin Damage Risks, and Protective Action*. Draft Report prepared by University of Wyoming for US Environmental Protection Agency, Washington, DC.

Doyle, J.K., S.R. Elliot, G.H. McClelland, and W.D. Schulze. 1991. *Valuing the Benefits of Groundwater Cleanup: Interim Report.* Report by the University of Colorado to the U.S. Environmental Protection Agency.

Doyle, James K., Gary H. McClelland, William D. Schulze, Steven R. Elliott, and Glenn W. Russell. 1991. "Protective Responses to Household Risk: A Case Study of Radon Mitigation," *Risk Analysis*, Vol. 11, No. 1, pp. 121-34.

Edwards, Steven F., and Glen D. Anderson. 1987. "Overlooked Biases in Contingent Valuation Surveys: Some Considerations," *Land Economics*, Vol. 63, No. 2, pp. 168-178.

Fischhoff, Baruch. 1990. *Managing Visibility at National Parks in the Southwest: Review and Critique of a Contingent Valuation Study*. Report prepared by Research Triangle Institute, NC, for US EPA.

Fischhoff, Baruch, and Lita Furby. 1986. *A Review and Critique of Tolley, Randall et al.: "Establishing and Valuing the Effects of Improved Visibility in the Eastern United States."* Report

to the U.S. EPA by Eugene Research Institute, Eugene, OR.

Fisher, Ann, Gary H. McClelland, and William D. Schulze. 1988. "Measures of Willingness to Pay Versus Willingness to Accept: Evidence, Explanations and Potential Reconciliation," In G. L. Peterson, B. L. Driver, and R. Gregory, eds., *Amenity Resource Valuation: Integrating Economics with Other Disciplines*, Venture, State College, PA.

Fisher, Ann, and Robert Raucher. 1984. "Intrinsic Benefits on Improved Water Quality: Conceptual and Empirical Perspectives," in V. Kerry Smith and Ann D. Witte, eds., *Advances in Applied Microeconomics*, Vol. 3, JAI Press, Greenwich, Conn.

Gegax, Doug, Shelby Gerking, William D. Schulze, and Donald Anderson. 1985. *Experimental Methods for Assessing Environmental Benefits, Vol. IV, Valuing Safety: Two Approaches.* Report to the U.S. EPA by the University of Wyoming.

Gegax, Doug, Shelby Gerking, and William D. Schulze. 1991. "Perceived Risk and the Marginal Value of Safety," *Review of Economics and Statistics*, Vol. 73, No. 4, pp. 589-96.

Gerking, Shelby, Menno de Haan, and William D. Schulze. 1988. "The Marginal Value of Job Safety: A Contingent Valuation Study," *Journal of Risk and Uncertainty*, Vol. 1, No. 2, pp. 185-99.

Gregory, Robin. 1986. "Interpreting Measures of Economic Loss: Evidence from Contingent Valuation and Experimental Studies," *Journal of Environmental Economics and Management*, Vol. 13, No. 14, pp. 325-337.

Gregory, Robin S., Baruch Fischhoff, Lita Furby, and Jack Knetsch, *et al.* 1985. *Measures of Consumer's Surplus: Interpreting the Disparity in Views*, prepared by University of Wyoming for US Environmental Protection Agency.

Gregory, Robin, and Lita Furby. 1987. "Auctions, Experiments, and Contingent Valuation," *Public Choice*, Vol. 55, No. 3, pp. 273-289.

Hammitt, James K. 1986. *Estimating Consumer Willingness to Pay to Reduce Food-Borne Risks*, R-3447, The RAND Corporation, Santa Monica, CA. Report to US EPA

Harrington, Winston, Alan Krupnick, and Walter Spofford. 1985. *The Benefits of Preventing an Outbreak of Giardiasis Due to Drinking Water Contamination*, prepared by Resources for the Future for US Environmental Protection Agency.

Harrison, David, Jr. 1984. *Research and Demonstration of Improved Methods for Carrying out Benefit-Cost Analyses of Individual Regulations*, Vols. I, II, III, & IV, prepared by Harvard University for US Environmental Protection Agency.

Jones, Carol A., and Yusen D. Sung. *1991. Valuation of Environmental Quality at Michigan Recreational Fishing Sites: Methodological Issues and Policy Applications*, prepared by University of Michigan, School of Natural Resources for US Environmental Protection Agency.

Kealy, Mary Jo, Mark Montgomery, and John F. Dovidio. 1990. "Reliability and Predictive Validity of Contingent Values: Do the Characteristics of the Good Matter?," *Journal of Environmental Economics and Management*, Vol. 19, pp. 244-263.

Kealy, Mary Jo, and Robert W. Turner. 1993. "A Test of the Equality of Close-Ended and Open-Ended Contingent Valuations," *American Journal of Agricultural Economics*, Vol. 75, No. 2, pp. 321-33.

Kneese, Allen V. 1984. *Measuring the Benefits of Clean Air and Water*. Resources for the Future, Washington, D.C.

Lazo, J. K., William D. Schulze, Gary H. McClelland, and James K Doyle. 1992. "Can Contingent Valuation Measure Nonuse Values?" *American Journal of Agricultural Economics*, December, pp. 1126-32.

Loehman, Edna T., and David Boldt. 1990. *Valuing Gains and Losses in Visibility and Health with Contingent Valuation*, prepared by Purdue University for US Environmental Protection Agency.

Magat, Wesley A., W. Kip Viscusi, and Joel Huber. 1988. "Paired Comparison and Contingent Valuation Approaches to Morbidity Risk Valuation," *Journal of Environmental Economics and Management*, Vol. 15, No. 4, pp. 395-411.

McClelland, Gary, William Schulze, and Edward Balistreri. 1994. *An Examination of Performance Testing Requirements for Contingent Valuation*. Draft report to U.S. EPA by the University of Colorado. Included as Appendix B of these comments.

McClelland, Gary H., William Schulze, and D. Coursey. 1987. *Improving Accuracy and Reducing Costs of Environmental Benefit Assessments*. Report prepared by University of Wyoming for US Environmental Protection Agency.

McClelland, Gary H., William Schulze, Donald Waldman, Julie Irwin, and David Schenk. 1990. "Sources of Error in Contingent Valuation." Paper prepared for U.S. EPA at the University of Colorado.

McClelland, Gary, William Schulze, Donald Waldman, and D. Jay Goodman. 1994a. *An Examination of the Proposed Scope Test using Market Data*. Draft report to U.S. EPA by the University of Colorado. Included as Appendix C of these comments.

McClelland, Gary H., William Schulze, Donald Waldman, Julie Irwin, David Schenk, Thomas Stewart, Leland Deck, and Mark Thayer. 1993. *Innovative Approaches for Valuing the Perceived Environmental Quality*: *Valuing Eastern Visibility: A Field Test of the Contingent Valuation Method*. Report for U.S. EPA from University of Colorado, Boulder, CO, September.

McClelland, Gary H., William D. Schulze, Jeffrey K. Lazo, Donald H. Waldman, James K. Doyle, Steven R. Elliott, and Julie R. Irwin. 1992. *Methods for Measuring Non-Use Values: A Contingent Valuation Study of Groundwater Cleanup*. Report to U.S. EPA from University of Colorado, Boulder, CO.

McConnell, Kenneth E., and I.E. Strand. 1994. *The Economic Value of Mid-and South Atlantic Sportfishing*. Report to the U.S. EPA by University of Maryland.

Mendelsohn, Robert. 1986. Controlling Outliers in Contingent Valuation Experiments, prepared by Yale University, School of Forestry for U.S. Environmental Protection Agency.

Meta Systems Inc. 1985. *A Methodological Approach to an Economic Analysis of the Beneficial Outcomes of Water Quality Improvements from Sewage Treatment Plant Upgrading and combined Sewer Overflow Controls*. Report prepared for US Environmental Protection Agency.

Milon, J. Walter. 1989. "Contingent Valuation Experiments for Strategic Behavior," *Journal of Environmental Economics and Management*, Vol. 17, pp. 293-308.

Mitchell, Robert C., and Richard T. Carson. 1981. *An Experiment in Determining Willingness to Pay for National Water Quality Improvements.* Draft report by Resources for the Future to USEPA (ORD).

Mitchell, Robert C., and Richard T. Carson. 1984. *A Contingent Valuation Estimate of National Freshwater Benefits: Technical Report to the U.S. Environmental Protection Agency*, Resources for the Future, Washington, D.C.

Mitchell, Robert C., and Richard T. Carson. 1986. *Valuing Drinking Water Risk Reductions using the Contingent Valuation Method: A Methodological Study of Risks from THM and Giardia*. Report prepared by for USEPA by Resources for the Future, Washington, DC.

Mitchell, Robert C., and Richard T. Carson. 1986. *The Use of Contingent Valuation Data for Benefit/Cost Analysis in Water Pollution Control*. Report prepared by Resources for the Future for U.S. Environmental Protection Agency.

Mitchell, Robert C., and Richard T. Carson. 1989. *Existence Values for Groundwater Protection*, prepared by Resources for the Future for US Environmental Protection Agency.

Peterson, Donald C., Robert D. Rowe, and W. Schulze. 1987. *Improving Accuracy and Reducing Costs of Environmental Benefit Assessments: Valuation of Visual Forest Damages from Ozone.* Report prepared by University of Colorado for U.S. Environmental Protection Agency.

Randall, A., J. P. Hoehn, and G. S. Tolley. 1982. "The Structure of Contingent Markets: Some Results of a Recent Experiment," *American Economic Review*.

Randall, Alan, and Glenn C. Bloomquist. 1985. *National Aggregate Benefits of Air and Water Pollution Control*—Volumes I & II, prepared by University of Kentucky for US Environmental Protection Agency.

Roberts, Marc J. 1974. *A Study of the Measurement and Distribution of Costs and Benefits of Water Pollution Control.* Final Report to U.S. EPA by Harvard University.

Rowe, Robert D., Ronald Dutton, and Lauraine Chestnut. 1985. *The Value of Ground Water Protection: Miami Case Study Design and Pretest.* Report to the U.S. EPA. Energy and Resource Consultants, Inc., Boulder, CO.

Rowe, Robert D., Lauraine G. Chestnut, and M. Skumanich. 1990. *Controlling Wintertime Visibility Impacts at the Grand Canyon National Park: Social and Economic Benefits Analysis*, prepared by RCG / Hagler, Bailly, Inc., for US EPA.

Rowe, Robert D., and Lauraine G. Chestnut. 1986. *Addendum to Oxidants and Asthmatics in Los Angeles: A Benefit Analysis.* Report prepared by EPA Office of Policy, Planning and Evaluation for U.S. Environmental Protection Agency.

Rowe, Robert D., Ralph C. D'Arge, and David S. Brookshire. 1980. "An Experiment on the Economic Value of Visibility," *Journal of Environmental Economics and Management*, Vol. 7, pp. 1-19.

Schulze, William D., R. R. Boyce, T. C. Brown, G. H. McClelland, and G. L. Peterson. 1992. "An Experimental Examination of Intrinsic Environmental Values," *American Economic Review*, Vol. 82, No. 5, May, pp. 1366-72.

Schulze, William D., Ronald G. Cummings, David S. Brookshire, Mark A. Thayer, Regan Whitworth, and Morteza Rahmatian. Undated. *Methods Development in Measuring Benefits of Environmental Improvements: Volume II, Experimental Approaches for Valuing Environmental Commodities*, report prepared by the University of Wyoming for USEPA.

Schulze, W. D., D.S. Brookshire, and E.G. Walter, and K. Kelley. 1981. *The Benefits of Preserving Visibility in the National Parklands of the Southwest: Volume 8 of Methods Development for Environmental Control Benefits Assessment*, US Environmental Protection Agency,

University of Wyoming, Resource & Environmental Economics Lab.

Schulze, W.D., D.S. Brookshire, and E.G. Walter, "The Economic Benefits of Preserving Visibility in the National Parklands of the Southwest," *Natural Resources Journal*, Vol. 23, pp. 149-73.

Schulze, William D., J. R. Irwin, D. J. Schenk, G. H. McClelland, T. Streard, L. Deck, and M. Thayer. 1990. "Urban Visibility: Some Experiments on the Contingent Valuation Method," in C. V. Mathai, ed., *Visibility and Fine Particles, Transactions of the Air and Waste Management Association*, Pittsburgh, PA, pp. 647-58.

Schulze, William, Gary McClelland, Ed Baliistreri, Rebecca Boyce, Michael Doane, Brian Hurd, and Roanld Sminauer. 1994. *An Evaluation of Public Preferences for Superfund Site Cleanup, Volume 2: Pilot Study*, Draft Report by the University of Colorado to USEPA, March.

Schulze, William, G. McClelland, D. Brookshire, and D. Coursey. 1985. *Improving Accuracy and Reducing Costs of Environmental Benefit Assessments: Vol. V, Experimental Approaches for Measuring the Value of Environmental Goods*, prepared by University of Colorado for US Environmental Protection Agency.

Schulze, William, G. McClelland, Brian Hurd, and Joy Smith. 1986. *Improving Accuracy and Reducing Costs of Environmental Benefits Assessments: Volume IV, A Case Study of a Hazardous Waste Site: Perspectives from Economics and Psychology*, Report by the University of Colorado to USEPA, May

Schulze, William, G. McClelland, D. Waldman, D. Schenk, and J. Irwin. 1990. *Valuing Visibility: A Field Test of the Contingent Valuation Method*, report prepared by University of Colorado for USEPA.

Schulze, William D., Gary H. McClelland, David J. Schenk, Julie R. Irwin, Steven R. Elliott, Rebecca R. Boyce, Thomas Stewart, Paul Slovic, Sarah Lichtenstein, Leland Deck, and Mark Thayer. 1993. *Improving Accuracy and Reducing Cost of Environmental Benefit Assessments: Field and Laboratory Experiments on the Reliability of the Contingent Valuation Method*. Report by University of Colorado to the U.S. Environmental Protection Agency, Washington, DC, September.

Shogren, Jason F. 1988. *Valuing Risk in Experimental Markets: Self-Protection, Self-Insurance, and Collective Action.* Report to U.S. EPA by Appalachian State University.

Smith, V. Kerry, and William H. Desvousges. 1986. "The Value of Avoiding a LULU: Hazardous Waste Disposal Sites," *Review of Economics and Statistics*, Vol 78, No. 2, pp. 293-99.

Smith, V. Kerry, and William H. Desvousges. 1987. "An Empirical Analysis of the Economic Value of Risk Changes," *Journal of Political Economy*, Vol. 95, pp. 89-114.

Smith, V. Kerry, and William H. Desvousges. 1990. "Risk Communication and the Value of Information: Radon as a Case Study," *Review of Economics and Statistics*, February, pp. 137-42.

Smith, V. Kerry, William H. Desvousges, and Ann Fisher. 1983. "Estimates of the Option Values for Water Quality Improvements," *Economic Letters*, Vol. 13, No. 1, pp. 81-86.

Smith, V. Kerry, and Ann Fisher. 1986. "A Comparison of Direct and Indirect Methods for Estimating Environmental Benefits." *American Journal of Agricultural Economics*, Vol. 68, pp. 280-90.

Smith, V. Kerry, and Raymond B. Palmquist. 1989. *The Value of Recreational Fishing on the Ablemarle and Palmico Estuaries.* Report prepared for the U.S. Environmental Protection Agency.

Smith, V. Kerry, and William H. Desvouges. 1990. "Risk Communication and the Value of Information: Radon as a Case Study." *Review of Economics and Statistics*, February, pp. 137-42.

Smith, V. Kerry, William H. Desvousges, and A. Myrick Freeman. 1985. *Valuing Changes in Hazardous Waste Risks: A Contingent Valuation Analysis* - Vol. I, II, III, prepared by Vanderbilt University; Research Triangle Institute for US Environmental Protection Agency (EARB).

Thayer, M. 1981. "Contingent Valuation Techniques for Assessing Environmental Impacts: Further Evidence," *Journal of Environmental Economics and Management*, Vol. 8, pp. 27-44.

Thayer, M., and W. Schulze. 1977. *Valuing Environmental Quality: A Contingent Substitution and Expenditure Approach*, University of Southern California.

Tolley, George S., *et al.* 1986. Results from 1984 Contingent Valuation Study of Visibility and Comparison with 1982 Results: Photos, Vehicle, Seasonality and Distribution. Report to U.S. EPA. University of Chicago.

Tolley, George, A., Randall, G. Bloomquist, and M. Brien, *et al.* 1986. *Establishing and Valuing the Effects of Improved Visibility in Eastern United States*, prepared by University of Chicago for US Environmental Protection Agency (OPA).

Tolley, George, and Lyndon Babcock. 1986. *Valuation of Reductions in Human Health Symptoms and Risks*—Vols. I, II, III, & IV, prepared by University of Chicago for US Environmental Protection Agency (EARB).

Tolley, George S., and Robert Fabian, eds. 1988. *The Economic Value of Visibility*. The Blackstone Co., Mt. Pleasant, MI.

van Ravenswaay, Eileen O. and John P. Hoehn. 1990. *The Impact of Health Risk on Food Demand: A Case Study of Alar and Apples*. Staff Paper No. 90-31. Department of Agricultural Economics, Michigan State University.

van Ravenswaay, Eileen O. and John P. Hoehn. 1991. *Consumer Perspectives on Food Safety Issues: The Case of Pesticide Residues in Fresh Produce*. Staff Paper No. 91-20. Department of Agricultural Economics, Michigan State University.

van Ravenswaay, Eileen O. and John P. Hoehn. 1991. *Consumer Willingness to Pay for Reducing Pesticide Residues in Fresh Produce*. Staff Paper No. 91-13. Department of Agricultural Economics, Michigan State University.

van Ravenswaay, Eileen O. and John P. Hoehn. 1992. "Consumers' Willingness to Pay for Risk Reduction When Risks are Ambiguous." Paper presented at the NE-165 Valuing Foods Safety and Nutrition Workshop. Alexandria, VA, June.

Viscusi, W. Kip. 1993. "The Value of Risks to Life and Health," *Journal of Economic Literature*, Vol 31, No. 4, December, pp. 1912-46.

Viscusi, W. Kip, and William Evans. 1991. "Estimation of State-Dependent Utility Functions using Survey Data," *Review of Economics and Statistics*, Vol. 73, No. 1, February, pp. 93-104.

Viscusi, W. Kip, and William Evans. 1991. "Utility-Based Valuations of Health," *American Journal of Agricultural Economics*, Vol. 73, No. 2, December, pp. 1422-7.

Viscusi, W. Kip, and William Evans. 1993. "Income Effects and the Value of Health," *Journal of Human Resources*, Vol. 28, No. 3, Summer, pp. 497-518.

Viscusi, W. Kip, Wesley A. Magat. 1987. *Learning about Risk: Consumer and Worker Responses to Hazard Information*. Harvard University Press, Cambridge, MA.

Viscusi, W. Kip, Wesley A. Magat. 1992. *Informational Approaches to Regulation*. Regulation of Economic Activity Series No. 19. MIT Press, Cambridge, MA.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. 1987. "An Investigation of the Rationality of Consumer Valuations of Multiple Health Risks," *Rand Journal of Economics*, Vol. 18, No. 4, Winter, pp. 465-79.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. 1988. "Paired Comparison and Contingent Valuation Approaches to Morbidity Risk Valuation," *Journal of Environmental*

Economics and Management, Vol. 15, No. 4, December, pp. 395-411.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. 1989. *Pricing Environmental Health Risks: Survey Assessments of Risk-Risk and Risk-Dollars Trade-Off*. Report to the U.S. Environmental Protection Agency. Duke University.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. 1991. Pricing Environmental Health Risks: Survey Assessments of Risk-Risk and Risk-Dollar Trade-Offs for Chronic Bronchitis," *Journal of Environmental Economics and Management*, Vol. 21, No. 1, July, pp. 32-51.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. Forthcoming. "The Death Risk Lottery Metric for Valuing Health Risks: Applications to Cancer and Nerve Disease," *Management Science*.

Violette, Daniel M., and Lauraine G. Chestnut. 1989. Valuing Risks: New Information on the Willingness to Pay for Changes in Fatal Risks, prepared by Energy and Resource Consultants, Inc. for US Environmental Protection Agency (OPA).

Walsh, Richard G., Donn M. Johnson, and John R. McKean. 1992. Benefit Transfer of Outdoor Recreation Demand Studies, Water Resources Research, Vol. 28, No. 3, pp. 707-713.

APPENDIX B. AN EXAMINATION OF PERFORMANCE TESTING REQUIREMENTS FOR CONTINGENT VALUATION

Appendix B is an unpublished report submitted to U.S. EPA by the University of Colorado as part of their on-going research on contingent valuation for the Agency. The research represents what we believe to be unique research on performance testing requirements for contingent valuation. Since this is a critical issue for the proposed regulations and the draft report is not available from any other source, it is reproduced in this Appendix in its entirety.

AN EXAMINATION OF PERFORMANCE TESTING REQUIREMENTS FOR CONTINGENT VALUATION

by

Gary McClelland *

William Schulze **

Edward Balistreri ***

Center for Economic Analysis
University of Colorado
Boulder, Colorado 80309

USEPA Cooperative Agreement #CR-821980:

MEASURING THE SUBJECTIVE BENEFITS AND COSTS OF ENVIRONMENTAL PROGRAMS

September 1, 1994

Project Officer

Dr. Alan Carlin
Office of Policy, Planning and Evaluation
U.S. Environmental Protection Agency
Washington, DC 20460

* Department of Psychology, University of Colorado, Boulder.

** Department of Agricultural, Resource and Managerial Economics, Cornell University, Ithaca NY.

*** Department of Economics, University of Colorado, Boulder.

1. Introduction

The recent report by the NOAA Panel on Contingent Valuation (Arrow et al., 1993) suggested that one approach for validating a study attempting to measure the magnitude of passive or non-use values is to examine if values are sensitive to the scope of the injury or environmental cleanup. This "scope test" has now been included in the regulations proposed by NOAA for conducting a contingent valuation study (*Federal Register,* Jan. 7, 1994). The scope test can be justified on two grounds. First, a central axiom of economics is that more should be preferred to less. Second, it has been suggested that non-use values are insensitive to scope because respondents value the "warm glow" of a charitable contribution (Andreoni, 1990) rather than the environmental commodity described in the CV instrument (Kahneman and Knetsch, 1992).

This paper explores the theoretical issues raised by the proposed test and uses both an experimental and a simulation approach to examine questions raised in the theoretical analysis. It is shown theoretically in Section 2 that the proposed scope test can substantially increase the sample size required in a CV study. This section also discusses the types of bias that might arise if each respondent to a CV survey is required to answer more than one valuation question in an effort to limit the sample size required. Section 3 then develops an experimental design to explore issues of sample size and bias. We use as the commodity for hypothetical valuation by respondents an insurance policy which protects against a known

financial loss which occurs with a known probability (e.g., 50% chance of a $40 loss). This commodity has some appropriate characteristics. First, the proposed NOAA regulations apply to environmental damages which arise from oil spills. CV surveys evaluating oil spill damages (Rowe et al., 1991; Carson, et al., 1992) have usually described the commodity as a program which will eliminate any chance of a described oil spill (loss) which would occur with a specified probability if there were no program. Thus, our commodity is the simplest that we can construct and yet shares some essential characteristics with the real world commodity of interest. Second, we have extensively studied this commodity over a range of probabilities from .01 to .9 and over a range of losses from $4 to $40 in laboratory economics experiments using an incentive-compatible Vickrey auction mechanism (McClelland and Schulze, 1991, Irwin, McClelland, Schulze, 1992, McClelland, Schulze, Coursey, 1993). Thus, we know how subjects value our commodity in real as opposed to hypothetical markets. We vary the scope of the experimental commodity used here by varying the probability of a $40 loss and observe how hypothetical bids for insurance vary in response. In Section 4 we present the results of the experiment and use the resulting data set for a series of statistical simulations to explore issues such as the effect of using a split sample, and of using alternative question formats (open-ended WTP, payment card, or dichotomous choice) on the relative sample sizes required to satisfy the proposed scope test. We also examine the constraint placed on the allowable changes in scope. Section 5 presents our concluding remarks.

2. A Theoretical Analysis of the Proposed Performance Test for Contingent Valuation

The recently proposed NOAA regulations, published in the *Federal Register,* involve some novel features governing the admissibility of valuation measures for damage assesment.1 Of particular interest is the requirement that an acceptable contingent valuation study must be able to demonstrate sensitivity to differences in options. This is to be accomplished by finding a statistically significant difference in the values of options A and B, where each option is evaluated by *different*

---

1 The exact language of the proposed regulations is as follows *(Federal Register,* Jan. 7, 1994, p. 1183):

"Scope test. Controlling for attitudinal, demographic, perceptual, and other differences across respondents, the trustee(s) shall demonstrate statistically that the aggregate WTP across all respondents for the prevention or restoration program increases (decreases) as the scope of the environmental insult is expanded (contracted). The scope of the environmental insult is characterized by the severity of the natural resource injuries and the level of effectiveness and timing of the restoration or prevention program. The demonstration shall be conducted through the use of split samples."

"Number of scenarios. The trustee(s) shall administer to split samples different survey instruments containing three variations of the scope of the environmental insult that respondents perceive as different unless the trustee(s) can provide a reasonable showing that the three-scenario test is infeasible due to considerations of cost or lack of plausibility of scenarios. Where three scenarios are feasible, the statistical test shall involve pairwise comparisons. In either case, the scenarios may vary along any of the margins of intensity, geography, and duration of damage and , for prevention of scenarios, the probability of an event occurring. The trustee(s) shall document the rationale for the selected variations of the scope of the environmental insult. In determining the descriptions to be used with the split samples, the trustee(s) shall use realistic injury scenarios and prevention or restoration programs that the respondents accept as credible. **"**

"Maximum amount of differente between scenarios. The trustee(s) shall develop scenarios for the total value test. Prior to the performance of the test, the trustee(s) shall demonstrate that not more than ninety-five percent of respondents in a pre-test or in focus groups indicate that there are meaningful value differences between the scenarios to be tested in any pairwise comparison. The demonstration shall be based on a minimum of sixty valid responses. The trustee(s) shall exclude from this demonstration any individuals who indicate in screening questions that they are not willing to pay anything for any size environmental cleanup or who would be willing to pay unrealistically large and invariant amounts for any size environmental cleanup."

samples of respondents. So as to preclude trivial tests where options A and B are substantially different, the proposed regulations require that when both options are presented to the *same* sample, no more than 95% of the respondents indicate that options A and B are different. (Note that the proposed regulations also require that a third option, C, be shown to be significantly different from option A using a third sample. Since pairwise comparisons are required we ignore this third option in our theoretical analysis.)

The proposed regulations for ensuring value sensitivity are interesting, but unusual. Such an untried combination of between- and within-respondent judgments, although well-intentioned, may have unanticipated and dramatic consequences. A search of the literature reveals a number of theoretical papers concerned with the statistical and methodological issues underlying the choice of between-respondent versus within-respondent designs (e.g., Erlebacher, 1977; Greenwald, 1976; Keren, 1993; Keren and Raaijmakers, 1988; Nickerson and McClelland, 1989; Rosenthal and Rubin, 1980, Vonesh, 1983). This literature shows that the precision of between- and within-respondent estimates of value differences will be different. In particular, Erlebacher (1 977), Rosenthal and Rubin (1 980) and Keren (1 993) provide statistical tests for determining whether designs yield significantly different results. A formal analysis of this issue can be constructed (following Keren, 1993) as follows:

Let WTP_{Ai} represent the amount that the i-th individual is willing to pay for option A. We can model this as

$$WTP_{Ai} = \mu + \alpha_i + \varepsilon_{Ai}$$

where μ is the true population mean for option A, α_i is true amount that the i-th individual deviates from the mean (i. e., $\sum \alpha_i$ = O), and ε_{Ai} represents errors of measurement. Similarly, the amount that the i-th individual is willing to pay for Option B is modeled as

$$WTP_{Bi} = \mu + \alpha_i + \beta + \gamma_i + \varepsilon_{Bi}$$

where β represents the increment, if any, for the value of the potentially more valuable option B over option A and γ_i represents the individual variation in this value as a deviation around β (i.e., $\sum \gamma_i = 0$).

There are two ways to estimate the value difference between options A and B. The first is within-respondent where willingness-to-pay is obtained for both options from each respondent. Then, an estimate of the difference is given by

$$\begin{aligned} D_i = WTP_{Bi} - WTP_{Ai} &= (\mu - \mu) + (\alpha_i - \alpha_i) + \beta + \gamma_i + (\varepsilon_{Bi} - \varepsilon_{Ai}) \\ &= \beta + \gamma_i + (\varepsilon_{Bi} - \varepsilon_{Ai}) \end{aligned}$$

for each individual. The expected value of the average is

$$\overline{D} = \beta$$

and the expected variance of this mean is given by

$$\sigma^2_{Within} = (\sigma^2_{\gamma} + \sigma^2_{\varepsilon_B} + \sigma^2_{\varepsilon_A}) / n$$

where *n* is the number of respondents.

The second way to estimate the value difference is between-respondent where willingness-to-pay is obtained for option A from one group of *n* respondents and for option B from a separate group of *n* respondents. Then,

$$\begin{aligned} D_i = WTP_{Bi} - WTP_{Ai'} &= (\mu - \mu) + (\alpha_i - \alpha_{i'}) + \beta + \gamma_i + (\varepsilon_{Bi} - \varepsilon_{Ai'}) \\ &= \beta + \gamma_i + (\alpha_i - \alpha_{i'}) + (\varepsilon_{Bi} - \varepsilon_{Ai'}) \end{aligned}$$

where i and i' represent respondents from the different groups or samples. Again, we have

$$\overline{D} = \beta$$

but now the expected variance of this mean is

$$\sigma^2_{Btwn} = (2\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_{\varepsilon_B} + \sigma^2_{\varepsilon_A}) / n$$

It is then easy to compare the expected variances, and hence the expected precision, of the two estimates by subtracting to get

$$\sigma^2_{Btwn} - \sigma^2_{Wthn} = 2\sigma^2_\alpha / n$$

This difference is necessarily always positive so the between-respondent estimate will always have greater variance than the within-respondent estimate. [2] In other words, the within-respondent estimate is necessarily more precise statistically and the difference in precision is a function of the individual variability in values for Option A. This is true even though the between respondent test uses *2n* respondents and the within respondent test uses *n* respondents. The greater the variability in those individual values, the greater the superiority of the precision of the within-respondent estimate. Unfortunately, since the σ^2_α is the variance

---

2 Note that WTP_{Ai} and WTP_{Bi} are necessarily correlated. Specifically, their correlation is

$$r_{AB} = \frac{\sigma^2_\alpha}{\sqrt{(\sigma^2_\alpha + \sigma^2_{\varepsilon_A})(\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_{\varepsilon_B})}}$$

The effect of the measurement error is to attenuate the maximum possible correlation which is given by

$$\frac{\sigma_\alpha}{\sqrt{\sigma^2_\alpha + \sigma^2_\gamma}}$$

Squaring this last expression gives the maximum possible proportion of variance shared by the two WTP responses. That is,

$$\frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\gamma}$$

the ratio of the variance in individual absolute levels to the sum of that variance and the variance in individual relative differences between options A and B. The higher this correlation, the greater the superiority of the within-respondent design.

in the true value of the public good being valued, and since people are likely to have very different values for environmental cleanup, this variance is likely to be very large, implying that satisfying between-respondent tests will require very large sample sizes.

The psychology literature also suggests that absolute judgments of any kind are usually difficult, but that relative comparisons are much easier and hence more accurate (Baird and Noms, 1978). Between-respondent judgments are inherently absolute while within-respondent judgments are probably relative.

For example, the psychology of judgment suggests a simple thought experiment. Very few people have an ability to identify absolute pitch, but many people have a very good ability to identify relative pitch. If we tried to demonstrate a difference between Tone A and Tone B using a between-respondent design in which respondents rated a single tone, it would require an enormous sample size to detect a statistically significant difference between the judgments from the two samples. However, in a within-respondent design in which respondents heard both tones, a significant difference would likely be detected using a relatively small sample. Note also that it would likely be extremely difficult for two tones even just a half-step different on the musical scale to pass the test of only 95% of the respondents detecting a difference.

Less fanciful than the pitch example, consider a thought experiment from consumer choice. Suppose that Options A and B corresponded, respectively, to a VCR without remote control and a VCR with remote

control. Suppose a manufacturer tried to determine the value of remote control to consumers by following the proposed procedure. In a survey, such a manufacturer might have trouble passing the within-respondent requirement that no more than 9S% of the respondents would value options A and B were differently. But let's suppose that at least 5% of the respondents didn't value B more that A. Now, in the next survey the manufacturer asks one sample of respondents how much they would be willing to pay for A, the VCR without remote, and asks a different sample of respondents how much they would be willing to pay for B, the VCR with remote. There is likely to be considerable individual difference in willingness to pay for the base VCR and there will also be considerable individual difference in the willingness to pay for the extra feature of remote control. The result is that it will require a very large sample to overcome these sources of error so that the true average value of the remote control feature can be estimated. A feasible sample size would probably not be able to find a statistically significant difference. The likely outcome is that the manufacturer would decide incorrectly that the remote control feature had no value.

Now consider a within-consumer survey in which respondents were asked how much they were willing to pay for a base VCR and then how much more they would be willing to pay to get the remote control feature. For some consumers remote control has no value, in fact, considering the difficulty of programming some VCRs it might even be considered a disadvantage by some consumers; such consumers would likely give $0 as the amount more they would be willing to pay. For others, remote control would be an advantage and they would be willing to pay some positive

amount to get it. The manufacturer would get a much better estimate of the true average value of remote control to its consumers by using a within-consumer design.

It should also be noted that the Federal government spends millions of dollars each year to collect within subject data over time on economic behavior. The Panel Survey of Income Dynamics, the National Longitudinal Survey of Labor Market Experience, and the Survey of Income and Program Participation are examples of such data sets. The express purpose of tracking household behavior over time, as opposed to the less expensive approach of looking only at cross-sectional data, is that within-respondent comparisons allow the researcher to control for heterogeneity that exists across individuals, but that is not explained by the observed explanatory variables.

Within-respondent designs do, however, require that each respondent answer more than one question. The advantages of a within-respondent design in terms of precision and statistical power would not be worthwhile if there were substantial sequence effects that would bias responses after the first one (Rosenthal and Rubin, 1980). In the case of contingent values, there are two plausible hypotheses for sequence effects. The first is “anchoring and adjustment”. A number of judgment studies have observed that when making a subsequent judgment in the same domain, respondents tend to underadjust and so appear to be “anchored” to their initial response (see Tversky and Kahneman, 1974, Wright and Anderson, 1989, and Northcraft and Neale, 1987). For example, in a contingent valuation study estimating oil spill damages, a respondent

might be asked to bid on a program to eliminate 10% of all oil spills. The anchoring-and adjustment-hypothesis would predict that the respondent would anchor to that initial response and so underadjust this value upward to estimate the bid for, say, a program to eliminate 50% of all oil spills.

The second plausible sequence effect is due to social desirability or compliance (Schuman and Presser, 1981). That is, respondents may believe, in a survey that has asked two questions, that the researchers expect them to value a program to eliminate 50% of oil spills a lot more than one that would eliminate only 10% of oil spills. Respondents may comply with this expectation by overstating their true value differences between the two programs in order to please the researcher. Social desirability might also induce respondents to bid too much for both programs, in which case there would be an absolute error instead of or in addition to a relative error. Survey researchers attempt to avoid such biases by assuring respondents of anonymity. However, the possibility of overadjustment remains.

Our experiment is designed to address a number of the theoretical issues raised in the discussion presented above. To deal with the issue of the relative sample sizes required, we use the data we collect as the basis of a series of statistical simulations. Further, since we can collect hypothetical bids for insurance policies at different probabilities of loss in any order, we can investigate the issue of bias. Figure 1 shows the possibilities for absolute error in our experimental design. The horizontal axis shows the probability of loss, while the vertical axis shows the dollar bid for insurance. We use a fixed $40 loss throughout the

experiment, so risk aversion should not noticeably affect bids; thus, bids should approximately equal expected value (McClelland, Schulze, and Coursey, 1993). Where p is the probability and L denotes the loss, expected value is shown as EV in Figure 1. Where EV is taken as the true value, it is well known that actual absolute bids for insurance are biased (See, Edwards, 1954, Kahneman and Tversky, 1979, and Machina, 1982). In particular, actual bids for insurance above p=.4 fall below EV, a phenomenon known as underweighting, while bids for insurance below p=.4 are above EV, a phenomenon known as overweighting. As shown by McClelland et al. (1 993), experience over repeated rounds with actual risk in the Vickrey auction institution reduce both overweighting and underweighting. Thus in Figure 1, in moving between point a and point b (obtaining bids for p=.5 and bids for p=.6, and assuming no relative error) values will lie below the EV line if our experiment conforms to existing evidence. This is the range of probabilities we use in our design. Alternatively, if we use a range of probabilities below p=.4, as shown by moving between points a' and b', values will likely fall above EV. Based on the previous experiments cited above, we expect to find absolute error in our experiment with bids equal to about 80% of expected value.

Ignoring absolute error, Figures 2 and 3 show how our experimental results might reflect the two sorts of relative bias discussed above. In Figure 2, if respondents first provide a bid for p=.5, point a, but underadjust in attempting to provide bids for higher probabilities such as for p=.6, a bid such as that shown by point b will be obtained. If the order is reversed, respondents would move from point a' to point b'. Figure 3

FIGURE 1.

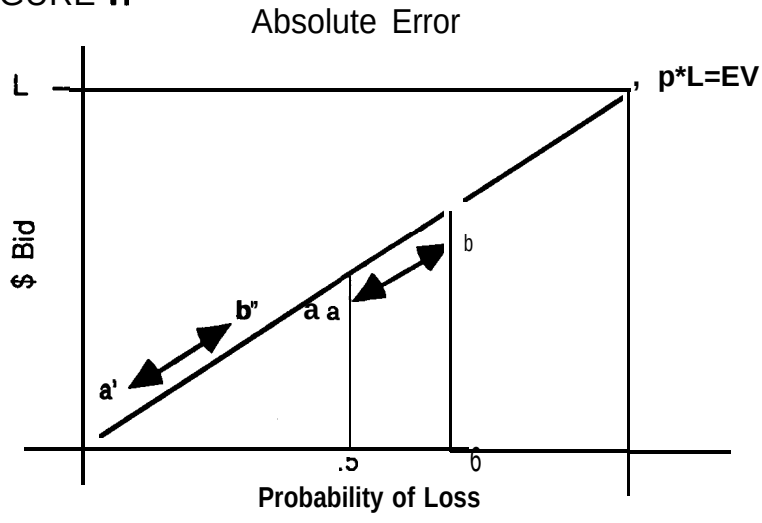


FIGURE 2.

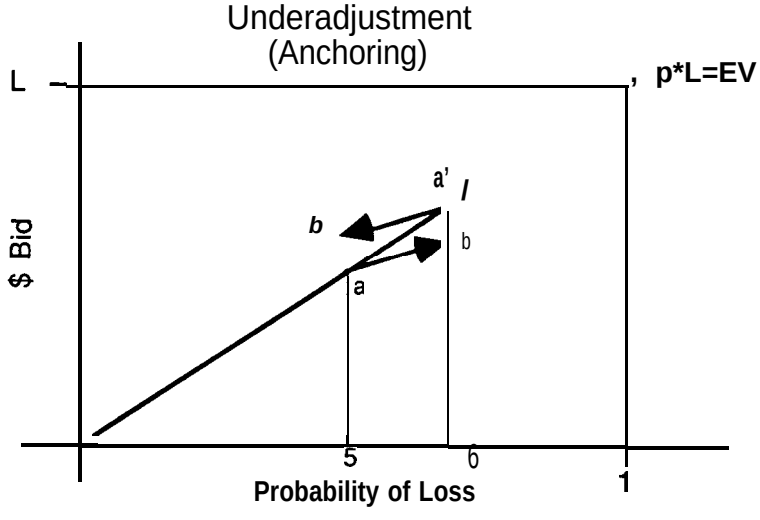
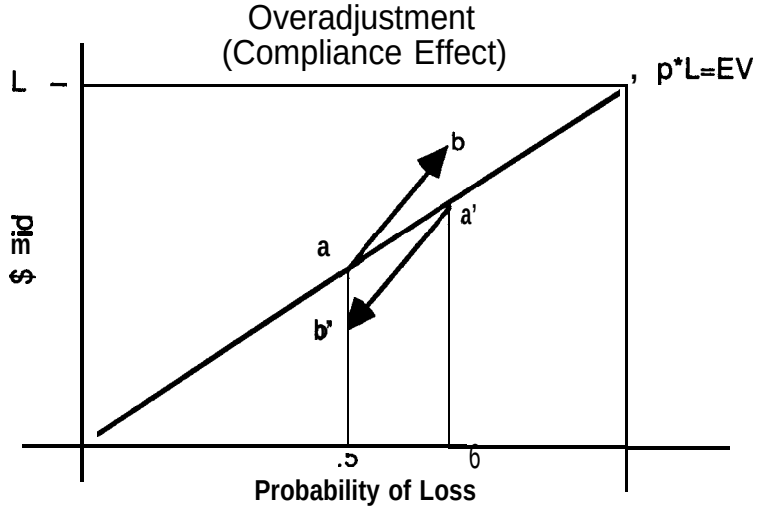


FIGURE 3.



shows the alternative hypothesis of overadjustment. Starting from point a, respondents, realizing that the experimenter wishes them to bid more for the higher probability insurance policy overadjust and bid too much at point b. A symmetrical argument leads to a move from point a' to b'.

Finally, the experimental design allows us to examine the issue of how the proposed regulations define the maximum allowed difference in scenarios -- that no more than 95% of respondents notice a difference in the value of the scenarios. Research in psychophysics provides an extensive literature on "just noticeable differences. " Many experiments have been conducted using a variety of stimuli (noise, taste, light, etc.) to explore the minimum detectable difference (in loudness, flavor, brightness, etc.). Where S is a measure of the intensity, the just detectable difference, AS, generally approximates Weber's Law,

$$AS = kS,$$

which implies that as the magnitude of the stimulus increases, the magnitude of the just noticeable difference increases proportionately. (For a discussion of Weber's Law see Chapter 3 of Baird and Noms, 1978.) Thus if an individual is asked to detect a difference in loudness between the two noises, the louder the noise, the greater the difference required. Obviously, Weber's law is not perfectly precise, so it is generally assumed that a random term (ε) is present; so

$$\Delta S = kS + \varepsilon.$$

It is also assumed that ε has a zero mean and a cumulative distribution $F(\varepsilon)$. in our experiment, the stimulus is the probability of loss, p, which we vary in small increments starting from p = .5 to p = .6 (or alternatively from p = .6 moving down to p = .5) and the response is the stated

willingness to pay for insurance. Thus, we can observe F, the fraction of respondents who have changed their bid from their bid for p = .S as the probability increases (or the fraction who change their bid from that for p = .6 as the probability decreases). Knowledge of F then allows us to calculate the maximum change in scope, AS (or Ap), which is allowed under the proposed regulations. Solving for ε using Weber's Law and substituting the result into F, the proposed regulation is met for any AS which satisfies:

$$F(AS - kS) \leq 0.95$$

Note that where p. is the starting probability, so AS = $p - p_0$ and S = p_0, the proposed rule becomes:

$$F(p - p_0 - kp_0) \leq 0.95.$$

For any of the distributions we use (cumulative normal, logit and Cauchy), the term kp_0 is absorbed in an estimated constant, allowing a unique estimate of the maximum value for AS=p - p. which would satisfy the proposed regulation.

3. The Experiment

The goal of the experiment was to generate data that could be used in a statistical analysis of within-respondent and between-respondent bid comparisons for a commodity. The commodity was chosen to be an insurance policy for a known financial hazard. The scope of the commodity can easily be controlled by varying the probability of the hazard event.

The experiment was designed to parallel the experiments of McClelland et al. (1993), but in a fully hypothetical framework. The context of the experiment asks the subjects to imagine that they have an initial balance of $50 and then asks how much they would be willing to pay to insure against a $40 loss. The loss occurs if a red poker chip is drawn randomly from a bag containing 100 red and white chips. The instructions emphasize a one time draw from a bag containing 100 chips, and explain the insurance commodity fully (See Appendix A for a copy of the instructions).

The experiment was administered to an undergraduate class at the University of Colorado, Boulder. The class consisted of 226 students. The survey was distributed to the class, and a verbal explanation of the instructions was given. To increase the saliency of the poker chip draw, a demonstration was staged in which one red chip was placed in a bag with 99 white chips. A student was asked to draw a chip. The student drew a white chip. The experimenter then explained what the consequences of this draw would have been had this been a real experiment (all the subjects would have kept their $50 balance). It was then further explained what would have happened if the red chip was drawn in a real experiment ($40 would be taken out of all the subjects' $50 balance).

Two probability of loss orders were developed that asked subjects four open-ended willingness-to-pay (WTP) questions for selected probabilities.[3] One survey started by asking for the willingness-to-pay

[3] A fifth question asked for insurance bids for 1 red chip out of 100. This question was only included to illustrate overweighting of low probabilities for a subsequent class discussion. This .

for the insurance given 50 red and 50 white chips. We refer to this as the 0.5-FIRST condition. The other survey first asked for a WTP given 60 red chips out of 100. We refer to this as the 0.6-FIRST condition. Table 1 shows the order in which the two surveys presented the WTP questions.

TABLE 1.

Question # (order)	Number of Red Chips (Probability of Loss x 100) 0.5-FIRST	0.6-FIRST
1	50	60
2	51	54
3	54	51
4	60	50

The two probability of loss orders (up versus down) were chosen to identify the possible relative biases illustrated in Figures 2 and 3.

An important part of experimental design is an analysis of the statistical power for detecting the anticipated effect. Judd and McClelland (1 989) ‘ give the following formula for estimating effect size for a two-group within-respondent design:

$$\eta^2 = \left[\frac{4\sigma^2}{(\mu_A - \mu_B)} + 1\right]^{-1}$$

Effect size tables are then consulted to determine the probability that the anticipated effect will be detected (i.e., there will be no Type II error) for given sample sizes. In this case, unlike in most contingent valuation studies, the range of bids for insurance will likely be constrained between O and the loss of $40. The maximum possible standard deviation of 20

last question was placed on a separate page and is used only to demonstrate the consistency of these data with those from prior experiments.

would occur if exactly half the respondents bid O and the other half bid $40. That is a most unlikely outcome so we will select a standard deviation of about half that as our guesstimate to be used to estimating an effect size; this guesstimate is also comparable to the standard deviation for p = 0.6 obtained in McClelland, et al. (1 993). If bids are approximately like those depicted in Figure 1, then the difference in means for the comparison between 50 and 60 chips will be about (60 - 50)($0.40) = $4.00. These yield an approximate anticipated effect size of η^2 = 0.04. The probability of detecting such an effect size with a sample of slightly more than 200 observations is about 0.8, which is generally considered to be an acceptable power. By using a commodity with a naturally constrained range of values we can explore methodological issues in an experimental setting with a fairly small sample size.

4. Results

Complete questionnaires were returned by 222 out of 226 students, 113 in the O. S-first group and 109 in the 0.6-first group.

In this section we first consider the relative power of between-respondent versus within-respondent comparisons to evaluate the estimated difference in value for insurance against a $40 loss with probability 0.5 versus 0.6. Then we address power implications for other ways (than open ended WTP) in which responses might be collected; namely, use of a payment card and dichotomous choice (referendum). Then we assess whether the estimated values are biased. Finally, we consider how these data would have fared in the within-respondent test designed

to ensure that alternatives for the scope test are reasonably close together.

4.1 Comparison of Statistical Power

Power of Mean Comparisons. The fundamental comparison is between the mean responses. In typical contingent valuation surveys, willingness-to-pay data are often highly skewed and/or contain extreme outliers. In such cases, before applying standard least-squares test statistics, it is necessary to remove outliers (using Windsorized or trimmed means or outlier indices) and/or transform the data (usually using Box-Cox transformations). Such transformations and outlier procedures were not necessary (or feasible) in this case because the response scale was effectively constrained at the top end and because typical responses were in the middle of this response scale, so statistically meaningful outliers were absent from the data.

Table 2 summarizes the results of the between- and within-respondent estimates of the difference in value for insurance at a probability of loss of 0.6 versus 0.5. The between-respondent difference, based on the first responses from the two order groups, is $5.48 and is reliably different from O, ($t(220) = 4.54$, $p < .0001$). There was no evidence for order effects in the within-subject estimates of the difference (the mean difference was $0.12 higher in the 0.5-first group, $t(220) = 1.3$, p = .1 9), so the two groups were combined. From the combined groups, the estimate of the difference is $4.98 and is reliably different from O ($t(221) = 10.76$, $p < .0001$).

Table 2.

Statistics for Between- and Within-Respondent Estimates of Value Differences

Comparison	$\overline{D}$	$\sigma^2_{\bar{D}}$	t	PRE	P	Minimum n	
						0.01	0.05
	Open-Ended Means						
Between	5.48	1.46	4.54	.085	<.0001	78	46
Within	4.98	.21	10.76	.34	<.0001	18	11
	Simulated Payment Card						
Between	2.70	.87	3.12	.042	.002	156	90
Within	2.48	.18	5.91	.136	<.0001	46	26

According to the statistical theory outlined above, the within-respondent estimate should be more precise (i.e., have a smaller variance) so long as the correlation between responses is positive. The correlation between responses for probabilities of 0.5 and 0.6 is 0.71 ($p <$ .0001), and indeed the variance of the within-respondent estimate is only 0.21 while the variance for the between-respondent estimate is 1.46. The lower variance of the estimate 'for the within-respondent comparison necessarily implies that it is a more powerful statistical test. A useful way to compare the differences in power is to ask at what minimum sample size the estimated difference would have been statistically significant. This is facilitated by computing a measure of the effect size that, unlike Student's t, does not depend on sample size. We use PRE, the proportional reduction in squared error (Judd & McClelland, 1989). The corresponding values for PRE are reported in Table 2. In this case, PRE has a beta distribution with parameters 0.5 and $df/2$. It is then easy to use tables of PRE (see Appendix C of Judd & McClelland, 1989) or numerical algorithms

to determine the minimal df and by implication the minimal sample size required for statistical significance.

The between-subject difference would have been significant at p = 0.01 with a total of 78 subjects (38 in each group). For p = 0.05, only 46 subjects (23 in each group) would have been required. The within-subject difference would have been significant at p = 0.01 and 0.05 with only 18 and 11 subjects, respectively. Thus, the ratio of the minimum number of subjects required for a significant scope difference for between versus within equals 4.3 for p = .01 and 4.2 for p = .05.

Simulated Payment Card. To reduce non-responses, contingent valuation studies often use a payment card on which respondents circle a dollar value which is the most they would be willing to pay. Experience indicates that these response scales substantially reduce missing data relative to open-ended willingness-to-pay questions (McClelland et al., 1992, and Mitchell and Carson 1989). We generated data for such a survey assuming that participants in this study would have followed instructions to circle the highest number they would be willing to pay for insurance. We used (as is typically done) a logarithmic response scale for the simulated data of $0, 1, 2, 4, 8, 16, 32. Thus, for example, if a respondent gave a bid for insurance of $15, we converted it for purposes of this analysis to $8, the highest response category below the actual bid.

Statistics for this analysis of simulated data are also reported in Table 2. Note that the estimated difference for both the between- and within-respondent comparisons ($2.70 and $2.48, respectively) are about half of

their values from the open-ended estimates, but they are still statistically greater than zero. Again, the variance of the estimate for the within-estimate is substantially less (0.18 versus 0.87), giving that comparison much more statistical power. The between-subject difference would have been significant at p = 0.01 with a total of 156 subjects (78 in each group). For p = 0.05, only 90 subjects (45 in each group) would have been required. Note that using a payment card has doubled the sample sizes required for between-subject scope tests. Also, the possible sequence effect is more substantial, with the O. S-first group having a difference that is $1.44 larger, but this is not statistically significant ($t(220) = 1.73$, p = .09), so both groups have been combined. The within-subject difference would have been significant at p = 0.01 and 0.05 with only 46 and 26 subjects, respectively. For the logarithmic response scale, the ratio of the minimum number of subjects required for a significant scope difference for between versus within equals 3.4 for p = .01 and 3.5 for p = .05.

Simulated Dichotomous Choice (Referendum) The preamble to the proposed regulations suggests a preference for the referendum method in which respondents simply indicate whether they would or would not vote for an option at a specified price. Different subgroups of respondents are given different prices. To compare the statistical power for this methodology, we again simulated data as if our respondents had participated in such a study. We only used their first responses and generated "votes" for insurance at assigned costs of $5, $10, $15, $25, and $35. Four to six prices are typically used in such studies (e.g. Carson et al., 1992). For example, if someone had bid $16, we generated "yes" votes

for $5, $10, and $15, and “no” votes for $25 and $35. Table 3 presents the resulting proportions of “yes” votes for each price.

Table 3.

Proportion “Yes” Votes as a Function of Price and Loss Probability

Price	p(Loss) = .5	p(Loss) = .6
$5	.946.	.963
$10	.823	.936
$15	.496	.716
$25	.159	.385
$35	.044	.055

A logistic analysis would be appropriate if different respondents had produced the votes at each price and probability of loss. This is not the case here so we cannot do a statistical analysis reporting levels of significance for these data. However, we estimate model parameters using logistic regression and then ask what sample sizes would need to be for statistical significance if the same proportions of “yes” votes had been obtained from independent subsamples. This strategy probably overestimates the statistical power of the logistic analysis given that these ten proportions actually come from only two groups of subjects. We estimated the parameters in the following model:

$$\Pr(\text{"}yes\text{"}) = \frac{e^{\beta_0+\beta_1 C+\beta_2 D}}{1+e^{\beta_0+\beta_1 C+\beta_2 D}}$$

where C = the cost and D = O or 1 if the probability of a loss is 0.5 or 0.6, respectively. We also estimated a second model which included a term $\beta_3 C*D$ to test for different slopes in the two order groups. Table 4 gives

the resulting parameter estimates and the minimum sample sizes needed for statistical significance assuming approximately equal numbers of independent respondents in each cost-probability group. Figure 4 displays the separate logistic functions for each group. The minimum sample sizes in Table 3 are those that would be required to detect a significant difference between the two logistic curves in Figure 4. The minimum required sample sizes of 255 (51 at each cost) and 150 (30 at each cost) for p = .01 and .05, respectively, are 3.3 times greater than the sample sizes required when the means were compared directly in a split sample comparison.

Table 4.

Parameter Estimates and Minimum Sample Sizes for Logistic Model of Simulated "Voting" Data [4]

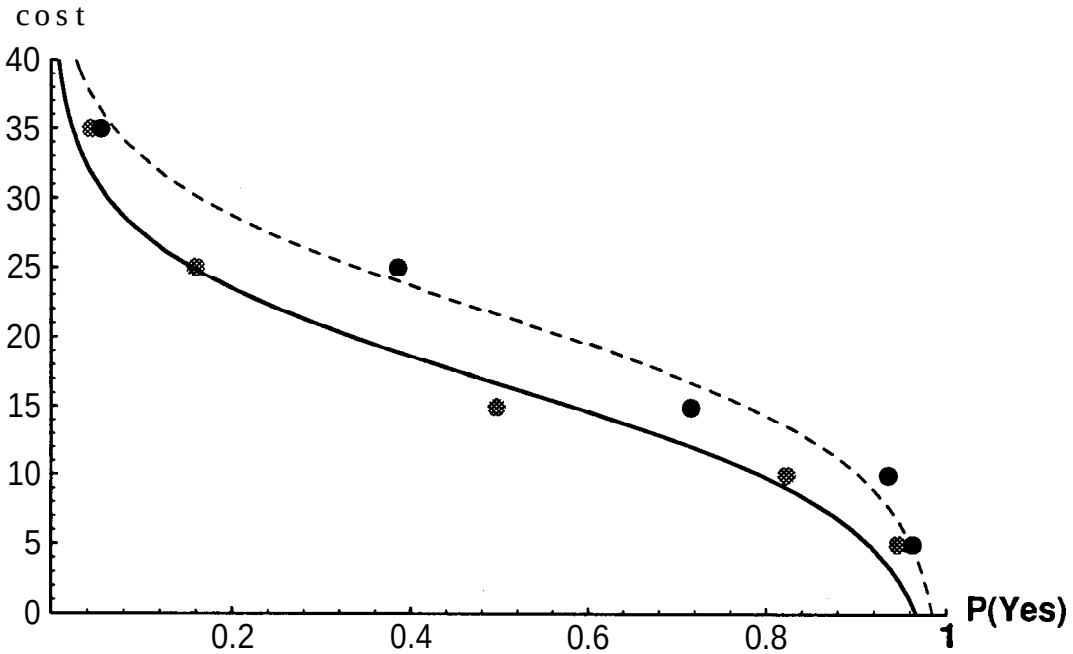
Source	Estimate	Minimum Sample	
		.01	.05
Intercept	4.27	——	——
Cost	-0.20	——	——
D (dummy)	0.96	255	150
Cost*D	0.01	36,000	21,200

Integrating the separate logistic functions in Figure 4 yields estimates of the mean bids for insurance. Doing so yields an estimated difference of $4.85, which is comparable to the estimate difference of $5.48 obtained by comparing the obtained means directly.

[4] The Intercept, Cost, and Dummy parameter estimates are from the first model which did not include the interaction term. Including the interaction term, which clearly is not important, slightly changes the estimate and interpretation of the other parameters.

Figure 4.

Simulated Vote Data and Fitted Logistic Functions for **P(Loss)** = 0.5 and 0.6.



Summary of Power Comparisons. Table 5 summarizes the relative statistical power of different response formats and comparison types. The most powerful comparison was the within-respondent comparison of mean bids so it is given a relative sample size of 1. The other entries in the relative sample size column indicate the relative increase in sample size necessary to have obtained statistical significance in this study. Thus, the 4.2 for the between-respondent comparison means that 4.2 times as many observations would be required for the between-respondent as the within-respondent comparison. The relative sample sizes are multiplicative so that a between-respondents study using the referendum format would require (4.2)(3.3) = 13.9 times as many observations as a within-respondent comparison of means.

Table 5.

Relative Sample Size Comparisons

Response Formats or Comparison Type	Relative Sample Size
Within-Respondents	1
Between-Respondents	4.2
Payment Card	2
Referendum Format	3.3

Using a between-respondent design instead of a within-respondent design has the largest effect on relative sample **size. The estimate of** 4.2 in Table 5 depends on characteristics of this particular study. It is therefore important to consider what values of relative sample size should be anticipated for other studies. This is most easily described if we use dimensionless effect size measures that do not depend on the number of observations. A difference between means in a between-respondent design is often tested with a t-test. It is well known that the correlation between a code (e.g., dummy or contrast) for the two groups and the responses is given by

$$r_b = \sqrt{\frac{t^2}{t^2 + 2(n-1)}}$$

where *n* is the number of observations in each group (assuming equal sample sizes in each group). This between-group correlation increases as the mean difference increases and decreases as the standard deviation increases. Also required to make the relative comparison is r_w, the correlation in a within-respondent design between responses to each option. Table 6 shows the relative sample size needed for a statistically significant difference at α = .05 as a function of r_b and r_w. The calculations underlying Table 6 assume equal numbers of respondents in

each between group, equal variances within each group, and normally-distributed errors with a common variance. As a point of reference, the between-group correlation in this experiment (for the 50 vs. 60 chip comparison) was approximately 0.3 and the within-respondent correlation was approximately 0.7. Interpolating in the row for $r_b = 0.3$ gives a relative sample size of about 4.5, which is slightly higher than the empirical value because of the error due to linear interpolation and because the assumptions were not perfectly satisfied.

Table 6.

Relative Sample Size for Different Effect Sizes

Between-Group Correlation (r_b)	Within-Respondent Correlation (r_w)				
	0.0	0.2	0.4	0.6	0.8
0.02	2.0	2.5	3.3	5.0	10.0
0.04	2.0	2.5	3.3	5.0	9.8
0.06	2.0	2.5	3.3	4.9	9.7
0.08	2.0	2.5	3.3	4.9	9.4
0.10	2.0	2.5	3.3	4.8	9.2
0.15	2.0	2.4	3.2	4.6	8.2
0.20	2.0	2.4	3.1	4.5	7.5
0.25	1.9	2.4	3.0	4.3	7.1
0.30	1.9	2.3	2.9	3.7	5.5
0.50	1.8	2.0	2.3	2.7	3.2

The empirical effect size of .3 obtained in this controlled classroom study based on a laboratory analog is much larger than would be obtained in typical contingent valuation studies. However, in general, the relative sample size does not depend much on the magnitude of the between-group correlation unless it is quite large. Instead, the relative sample size depends more on the magnitude of the within-respondent correlation. [n

general, the closer the options the higher the correlation will be. In this study, the correlation between bids for insurance at loss probabilities of 0.50 and 0.51 was .93, while the correlation for bids at probabilities of 0.50 and 0.60 was about 0.7. Ironically, the restriction against trivial scope tests means options will be closer together and so the superiority of within-respondent designs will be greater — but those are the specific designs that are ruled out for scope tests by the proposed regulations.

Note that even when there is no correlation between bids by the same respondents, there is still an advantage for within-respondent designs. This is because two data points (which would be independent due to the lack of correlation) are obtained from each respondent, thereby reducing the required sample size by 50%. Between-respondent designs would be superior in terms of required sample size only if the within-respondent correlation were negative. This might occur if the two options were incompatible or competing alternatives. For example, there would likely be a negative within-respondent correlation between responses to questions about how much money one was likely to contribute to (a) the Republican Party and (b) the Democrat Party.

The table for α = .01 comparable to Table 6 is very similar. All the relative sample sizes for α = .01 are the same or higher, but in most cases the differences are not large. For example, for r_b = 0.15 and r_w = 0.6, the relative sample size is 4.6 at α = .05 and 4.7 at α = .01. Thus, the values in Table 6 are conservative estimates of the increased sample size required for more restrictive levels of significance. In practice, required sample sizes are often determined by estimating anticipated effect sizes

(equivalent to guessing values of the between-group and within-respondent correlations in this case) and then calculating the sample size that would yield a given statistical power (often 0.80) of detecting the anticipated effect size. Calculating required sample sizes in this manner would slightly increase the relative sample size estimates in Table 6. For example, for r_b = 0.15 and r_w = 0.6, the relative sample size is 4.9 to achieve statistical power of 0.8. Again, the relative sample sizes in Table 6 are underestimates.

4.2 Accuracy

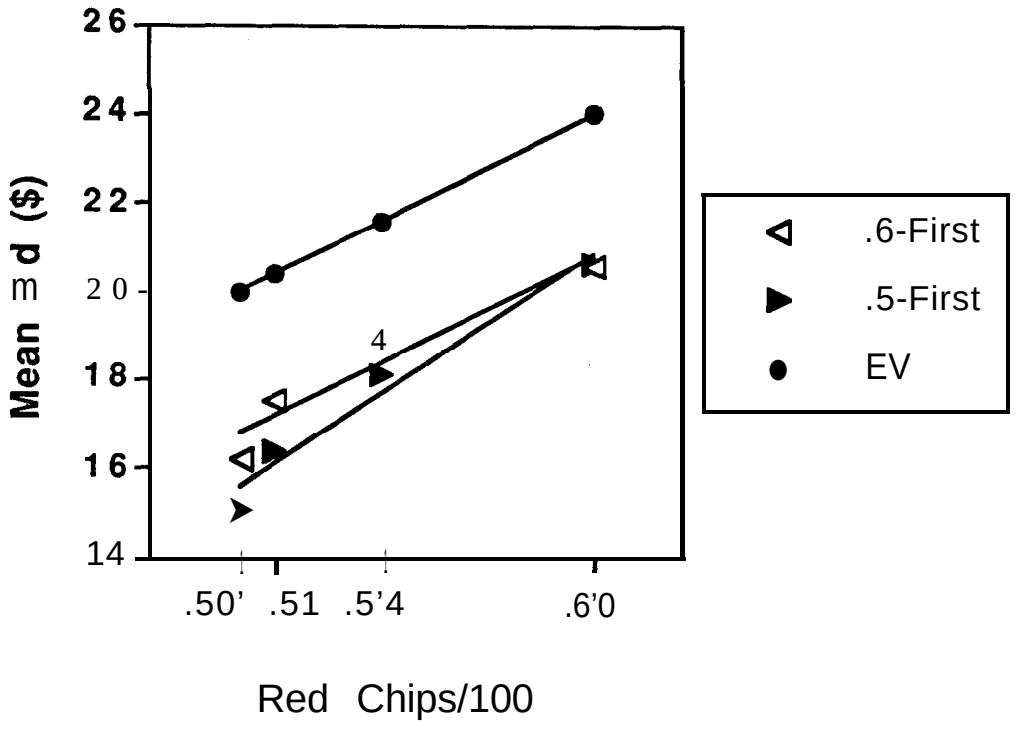
The superiority of within-subject designs would be moot if such designs introduced substantial bias. In this section we examine the accuracy of the estimates in terms of (a) whether the initial question biases subsequent responses (i. e., sequence effects), and (b) whether and in what pattern responses deviate from expected value. For greater statistical power we use responses from bids at all four probabilities (.50, .51, .54, .60). The data were analyzed by computing a separate regression equation for each respondent. The mean-deviated probability (minus the mean probability of .5375) was used as the predictor so that the intercept would equal each respondent's mean bid. If respondents had bid according to expected value then the resulting equation for each person would be

$$Bid = 21.50 + 0.40P'$$

where P' = Probability - 0.5375. Figure 5 displays the means for each group at each probability with the best-fitting regression line for each group.

Figure 5.

Mean Bids for Each Group at Each Probability
with Best-Fitting Lines for Each Group



Sequence Effects. There were no statistically reliable sequence effects in the prior analysis of comparisons between probabilities of 0.50 and 0.60. Consistent with those results, there are no significant differences between the best-fitting lines in Figure 5. The two slopes ($0. S3 for the O. S-first group and $0.40 for the 0.6-first group) are not reliably different ($t(220)$ = 1.48, p = *O.* 14) and the two intercepts or mean bids ($1 7.55 for the 0.5-first group and $18.31 for the 0.6-first group) are also not reliably different ($t(220)$ = 0.69, P = 0.49). In summary, there is no indication in any analysis that having respondents start either at the higher probability and then working down or at the lower probability and then working up had any effect on responses.

Comparison to Expected Value. Given that there are no statistical differences between the regression lines for each group, it is reasonable to combine them into a single model:

$$Bid = \mathbf{17.92 + 0.47P'}$$

The intercept and slope are reliably different from zero (respectively, $t(221) = 32.8$, $p < .0001$ and $t(221) = 10.8$, $p < .0001$). The more interesting question is whether these differ from the regression model expected if subjects were basing their bids on expected values. The intercept (mean bid) is reliably too low by \$3.58 ($t(221) = -6.55$, $p < 0.0001$), but the average slope of 0.47 is not reliably different from the expected slope of 0.40 ($t(221) = 1.52$, $p = 0.13$). Thus, of the three possible bias models outlined in Section 2, the results are most consistent with the absolute error model of Figure 1. In other words, consistent with the psychology of judgment, absolute judgments are difficult but relative judgments are relatively easier.

4.3 Non-Trivial Scope Test

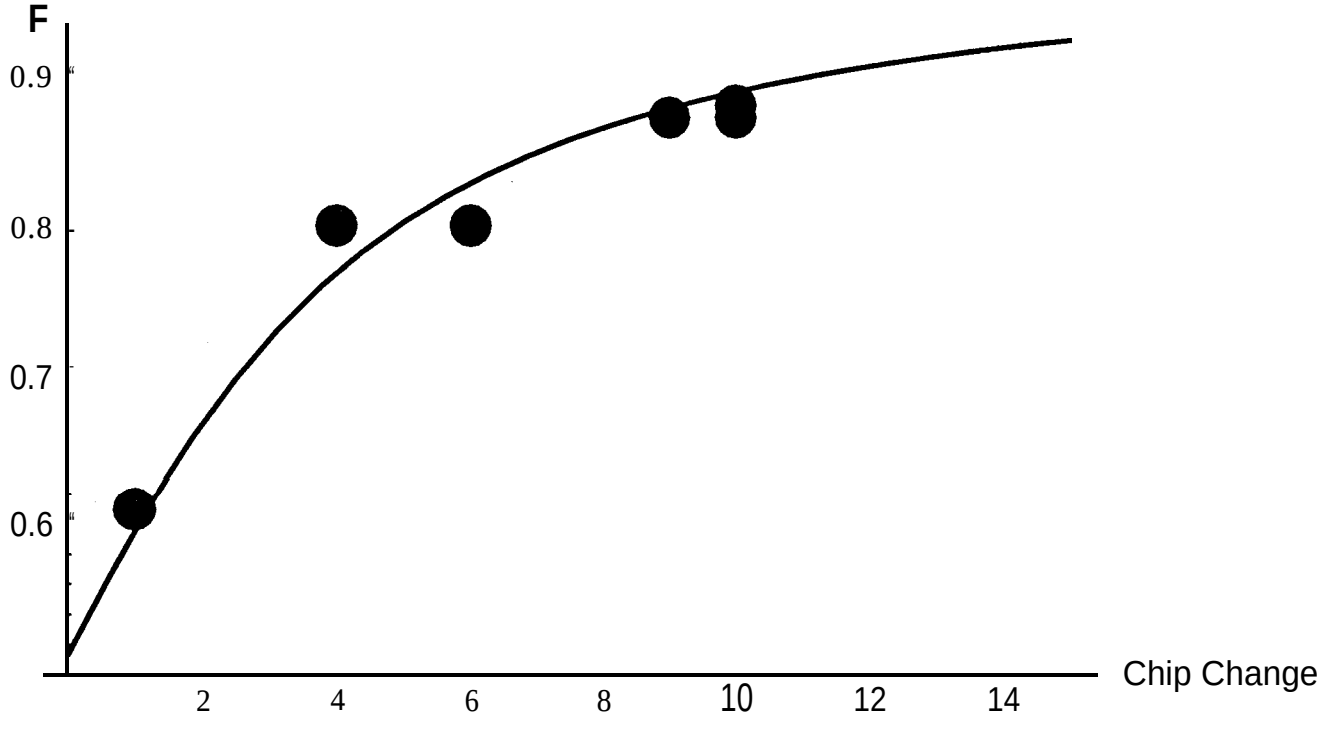
In order to preclude scope tests which are trivially satisfied by using greatly different options, the proposed regulations require that the options should not be so different that more than 95% of respondents state a value difference. It is interesting to ask how the present data would have fared with such a restriction. Figure 6 shows the percentage of respondents not changing their initial bid as a function of the distance (in chips or probability) from the initial bid; there were no reliable order effects so the data from both groups are combined in a single graph. These data easily pass the restriction against trivial scope tests because even for the maximum change in probability (0.1 0), only about 88% of the

respondents had changed their bids (regardless of whether they started at O .50 or 0.60).

To estimate how large the difference in chips would have been for 95% of respondents to have changed their bids, it is necessary to fit a function to these data and then extrapolate. Extrapolated predictions beyond the range of the data, especially those based on non-linear functions, should be accepted with extreme caution. We do so here just to derive an approximate estimate of how far apart the options in this study might have been and still not violated the restriction on the scope test. Fitting cumulative normal and logistic distribution functions to these data suggested that the data had thicker tails so the cumulative Cauchy distribution, with its thick tails, was tried. The best-fitting Cauchy function is depicted in Figure 6. Extrapolating this function suggests that for a probability change of about 0.23, approximately 95% of the respondents would have changed their bids. Thus, about a ±40% change from the base probability of loss is allowed by the proposed restriction on the change in scope.

Figure 6.

Percentage of Respondents Changing Initial Bid
as a Function of Changesin Probability of Loss
with Best-Fitting Cumulative **Cauchy** Function



5. Concluding Remarks

The manner in which the scope test is structured in the proposed regulations seems to be based on two assumptions. First, absolute values are unbiased. In other words, respondents are assumed to be capable of constructing values without a relative standard for comparison. Second, sequence effects are assumed to bias second responses so that relative values are likely to be biased. Consistent with the literature on the psychology of judgment, we find the pattern of bias in our experiment to

be precisely the opposite of that which underlies the proposed regulations. Relative values are relatively unbiased when compared to absolute values. It is important to note that one experiment provides insufficient evidence upon which to base firm conclusions. However, the existing evidence on bias does not provide strong support for the design of the scope test as currently proposed. For example, social desirability bias appears to become a major problem only for highly sensitive questions such as those on abortion, use of birth control, high risk behaviors for contracting AIDS, cheating on taxes, etc. (See for example Volumes I and II of Turner and Martin, 1984.) Thus, the benefits of the current specification of the scope test which requires a split sample are dubious.

The proposed regulations may impose two costs. First, if absolute values are biased, the prohibition against using relative values (which may be more accurate), may decrease the reliability of the estimates of non-use values obtained through contingent valuation. Second, the requirement of using a split sample for conducting the scope test greatly increases the costs of the test in terms of the required sample size. The sample size issue is further exacerbated by the requirement that a CV study must conduct not one, but two scope tests (using three scenarios presented to split samples). This requirement imposes another 50% increase in sample size. These increases in costs may be further magnified by the strong suggestion by NOAA to use the referendum approach, which in our analysis further increase the sample size required.

We cannot conclude that the proposed scope test is incorrect. However, the proposed test does not rest strongly on existing evidence in the

published literature on bias, nor is it consistent with the research presented here which finds absolute rather than relative error. Thus, the proposed structure of the test is based on speculation. This speculation, however, may come at a very high cost in increased sample size for conducting CV studies.

References

Andreoni, J., (1990) “Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving,” The Economic Journal, 100, 401.

Arrow, K, R. Solow, P.R. Portney, E.E. Learner, R. Radner and H. Schuman, (1993) “Report of the NOAA Panel on Contingent Valuation,” National Oceanic and Atmospheric Administration, U.S. Department of

Baird, J. C., and E. Noms, (1979) *Fundamentals of Scaling and Psychophysics, New* York: Wiley, 25-47.

Carson, R. T., R.C. Mitchell, W.M. Hanemann, R.J. Kopp, S. Presser, and P.A. Ruud, (1992) “A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill,” A Report to the Attorney General of the State of Alaska.

Edwards, W., (1954) “The Theory of Decision Making,” *Psychological Bulletin, 41,* 380-417.

Erlebacher, A., (1977) “Design and analysis of experiments contrasting the within- and between-subjects manipulations of the independent variable,” *Psychological Bulletin,* 84, 212-219.

Federal Register, Jan. 7, 1994, Part 11, Department of Commerce, National Oceanic and Atmospheric Administration, 15 CFR Part 990, Natural Resource Damage Assessments; Proposed Rules, 1139-1184.

Greenwald, A. G., (1976) “Within-subjects design: To use or not to use,” *Psychological Bulletin,* 83, 314-320.

Irwin, J. R., G.H. McClelland, and W.H. Schulze, (1992) “Hypothetical and Real Consequences in Experimental Auctions for Insurance Against Low-Probability Risks,” *Journal of Behavioral Decision Making,* 5, 107-116.

Judd, C. M., and G.H. McClelland, (1989) Data Analysis, San Diego: Harcourt Brace Jovanovich.

Kahneman, D., and J.L. Knetsch, (1992) Valuing Public Goods: The Purchase of Moral Satisfaction, " *Journal of Environmental Economics and Management,* 22, 57-70.

Kahneman, D. and A. Tversky, (1979) "Prospect Theory: An Analysis of Decision Under Risk," *Econometrics,* 47, 263-291.

Keren, G., (1993) "Between- or within-subjects design: A methodological dilemma," in G. Keren & C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues.* Hillsdale, NJ: Erlbaum.

Keren, G., and J.G.W. Raaijmakers, (1988) "On between-subjects versus within-subjects comparisons in testing utility theory, " *Organizational Behavior and Human Decision Processes,* 41, 233-291.

Machina, M. J., (1982) "Expected Utility Analysis without the Independence Axiom," *Econometrics,* 50, 227-323.

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

McClelland, G. H., and W, Schulze, (1991) "The Disparity Between Willingness-to-Pay and Willingness-to-Accept as a Framing Effect," in D.R. Brown and J.E. K. Smith (eds.), *Frontiers of Mathematical Psychology: Essays in Honor of Clyde Coombs,* 166-192.

McClelland, G. H., W. Schulze, and D.L. Coursey, (1993) "Insurance for Low-Probability Hazards: A Bimodal Response to Unlikely Events," Journal of Risk and Uncertainty, 7, 95-116.

McClelland, G. H., W. Schulze, J. Laze, D. Waldman, J. Doyle, S. Elliott, and J. Irwin, (1992) "Methods for Measuring Non-use Values: A Contingent Valuation Study of Groundwater Cleanup," U.S. Environmental Protection Agency Cooperative Agreement #CR-815183.

Nickerson, C. A., and G.H. McClelland, (1989) "Across-persons vs. within-persons tests of expectancy-value models: A methodological note," *Journal of Behavioral Decision Making,* 2, 261-270.

Northcraft, G. B., and M.A. Neale, (1987) "Experts, amateurs, and real-estate: An anchoring-and-adjustment perspective on property

pricing decisions," *Organizational Behavior and Human Decision Processes,* 39, 84-97.

Rosenthal, R., and Rubin, D., (1980) "Comparing within- and between-subjects studies," *Sociological Methods and Research,* 9, 127-136.

Rowe, R. D., W.D. Schulze, W.D. Shaw, D. Schenk, and L.G. Chestnut, (1991) "Contingent Valuation of Natural Resource Damage Due to the Nestucca Oil Spill, Final Report," prepared for State of Washington-Department of Wildlife, British Columbia-Ministry of Environment, and Environment Canada.

Schuman, H., and S. Presser, (1981) "Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context, San Diego: Harcourt Brace Jovanovich.

Turner, C. and E. Martin, eds., (1984) Surveying Subjective Phenomenon, New York: Sage Publishers.

Tversky, A., and D. Kahneman, (1974) "Judgment under uncertainty: Heuristics and biases," *Science,* 185, 1124-1130.

Vonesh, E.J. (1983) "Efficiency of Repeated Measures Designs Versus Completely Randomized Designs Based on Multiple Comparisons," *Communications in Statistical Theory and Methods,* 12, 289-302.

Wright, W. F., and U. Anderson, (1989) "Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assess merit," *Organizational Behavior and Human Decision Processes,* 44, 68-82.

Appendix A

Experimental Instructions

UP

Name_____
Student ID_____

Decision Making Under Uncertainty

This is a hypothetical experiment in the economics of decision making under uncertainty. We would like to know how much you would pay for an insurance policy to prevent the chance of a financial loss. Please read the following scenario carefully and do not hesitate to raise your hand if you have a question.

Imagine that you are given a starting balance of $50 for the experiment. (Experiments like this have been conducted for real at C.U.). Any money left at the end of the experiment is yours to keep. Further, imagine that there is a bag full of one hundred (100) poker chips: 50 red ones and 50 white ones. A chip is going to be picked randomly from the bag by a student in the class. If a white chip is drawn then you will keep your $50 and you owe nothing. If a red chip is drawn, however, you will have to pay $40; That is the loss of $40 will be deducted from your balance.

Rather than taking the chance of the $40 loss, you have the option of purchasing an insurance policy. If you buy the insurance policy then you will not owe the $40 in the event that a red chip is drawn. But, you will have to pay the experimenter, out of your balance, for the insurance policy before the chip is drawn.

We would like you to write down the most that you would pay for the insurance against the $40 loss for one draw from the bag. Although this experiment is hypothetical, please think about the problem carefully as if you really were facing this $40 loss if a red chip is drawn.

Given 50 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and_____cents.

Now imagine that instead of 50 whites and 50 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 51 red chips and 49 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 51 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____dollars and _____ cents.

Now imagine that instead of 50 whites and 50 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 54 red chips and 46 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 54 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

Now imagine that instead of 50 whites and 50 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 60 red chips and 40 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 60 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

Now imagine that instead of 50 whites and 50 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 1 red chip and 99 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 1 red chip out of 100, the most that 1 would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____cents.

DOWN

Name _____
Student ID _____

Decision Making Under Uncertainty

This is a hypothetical experiment in the economics of decision making under uncertainty. We would like to know how much you would pay for an insurance policy to prevent the chance of a financial loss. Please read the following scenario carefully and do not hesitate to raise your hand if you have a question.

Imagine that you are given a starting balance of $50 for the experiment. (Experiments like this have been conducted for real at C.U.). Any money left at the end of the experiment is yours to keep. Further, imagine that there is a bag full of one hundred (100) poker chips: 60 red ones and 40 white ones. A chip is going to be picked randomly from the bag by a student in the class. If a white chip is drawn then you will keep your $50 and you owe nothing. If a red chip is drawn, however, you will have to pay $40. That is the loss of $40 will be deducted from your balance.

Rather than taking the chance of the $40 loss, you have the option of purchasing an insurance policy. If you buy the insurance policy then you will not owe the $40 in the event that a red chip is drawn. But, you will have to pay the experimenter, out of your balance, for the insurance policy before the chip is drawn.

We would like you to write down the most that you would pay for the insurance against the $40 loss for one draw from the bag. Although this experiment is hypothetical, please think about the problem carefully as if you really were facing this $40 loss if a red chip is drawn.

Given 60 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

Now imagine that instead of 60 whites and 40 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 54 red chips and 46 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 54 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

Now imagine that instead of 60 whites and 40 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 51 red chips and 49 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 51 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____cents.

Now imagine that instead of 60 whites and 40 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 50 red chips and 50 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 50 red chips out of 100, the most that I would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

Now imagine that instead of 60 whites and 40 reds, a different number of red and white chips is placed in the bag. In this new situation you still have your $50 starting balance. The experimenter now places 1 red chip and 99 white chips in the bag. Carefully consider how much you would now pay for the insurance.

Given 1 red chip out of 100, the most that 1 would pay to prevent the chance of the $40.00 loss if a red chip is drawn is:

_____ dollars and _____ cents.

APPENDIX C. AN EXAMINATION OF THE PROPOSED SCOPE TEST USING MARKET DATA

Appendix C is an unpublished report submitted to U.S. EPA by the University of Colorado as part of their on-going research on contingent valuation for the Agency. The research represents what we believe to be unique research on performance testing requirements for contingent valuation. Since this is a critical issue for the proposed regulations and the draft report is not available from any sother source, it is reproduced in this Appendix in its entirety.

AN EXAMINATION OF THE PROPOSED SCOPE TEST USING MARKET DATA

by

Gary McClelland *

William Schulze **

Donald Waldman ***

D. Jay Goodman ***

Center for Economic Analysis
University of Colorado
Boulder, Colorado 80309

USEPA Cooperative Agreement #CR-821980:

MEASURING THE SUBJECTIVE BENEFITS AND COSTS OF ENVIRONMENTAL PROGRAMS

September 1, 1994

Project Officer

Dr. Alan Carlin
Office of Policy, Planning and Evaluation
U.S. Environmental Protection Agency
Washington, DC 20460

* Department of Psychology, University of Colorado, Boulder.

** Department of Agricultural, Resource and Managerial Economics, Cornell University, Ithaca NY.

*** Department of Economics, University of Colorado, Boulder.

AN EXAMINATION OF THE PROPOSED SCOPE TEST USING MARKET DATA

The proposed NOAA regulations on contingent valuation surveys (*Federal Register, Jan. 7, 1994)* include a "scope" test, which states that the researcher:

> "... shall demonstrate statistically that the aggregate [willingness to pay] WTP across all respondents for the prevention or restoration program increases (decreases) as the scope of the environmental insult is expanded (contracted) . .. The demonstration shall be conducted through the use of split samples. "

Additionally, the researcher:

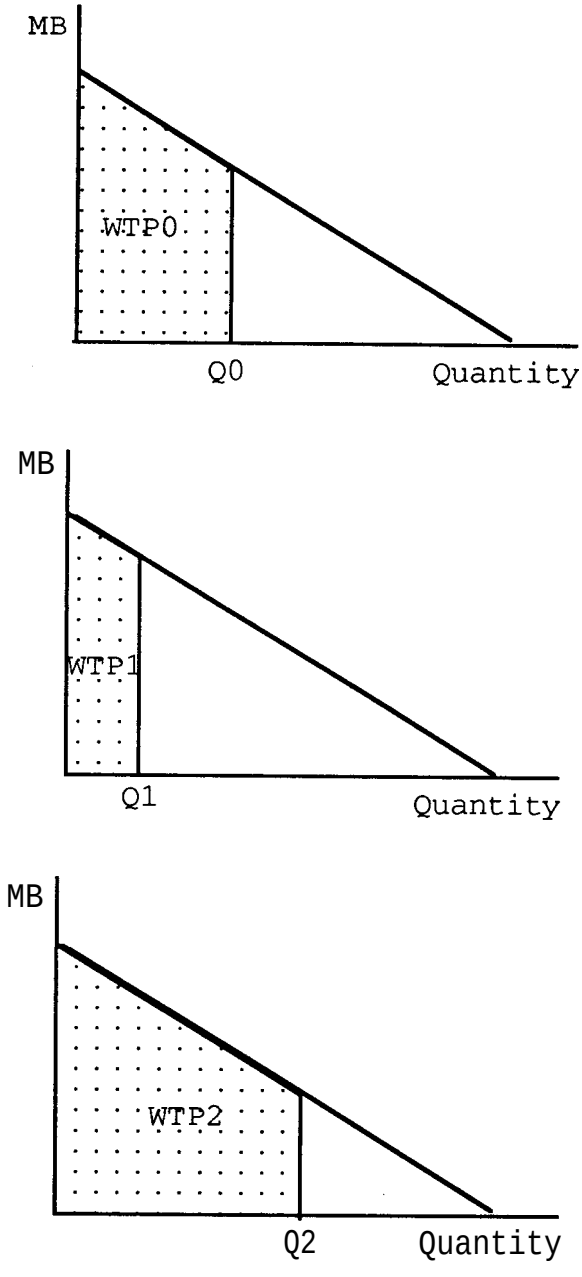
> "... shall administer to split samples different survey instruments containing three variation of the scope of the environmental insult that respondents perceive as different. " [emphasis added]

In order to fulfill these requirements, individuals in three separate samples must be asked their willingness to pay for different quantities of the environmental good. The NOAA proposal states that the researcher must show that the difference in the value of the environmental good is large enough that statistically significant differences in WTP can be detected between the samples. This test must be conducted for both increases and decreases in the quantity of the good.

With reference to Figure 1, which shows marginal benefits (MB) on the vertical axis and quantity (q) on the horizontal axis, the requirement would be to demonstrate that a smaller quantity q_1 offered to one sample results in a significantly lower WTP than that for the quantity q_0 offered to another sample ($WTP_1 < WTP_0$). Likewise, a greater quantity q_2 must be

shown to result in a significantly greater WTP than that associated with the quantity $(WTP_2 > WTP_0)$.

Figure 1 - Two-Way Scope Test of Significant Change in WTP



Since WTP for market goods can be easily estimated using demand data, this test can be done for any market good to estimate the change in

quantity necessary to produce a significant change in WTP. Extrapolating to non-market goods, the likely sample size requirements for contingent valuation studies under the proposed regulations can be estimated.

To test the implied hypotheses, a simple formula is developed for determining WTP in each case. Assume a linear demand curve, $Q = a + bP + e$, where e is normally distributed. The formula for WTP can be made to depend only on quantity and the estimated demand parameters $\hat{\alpha}$ and $\hat{\beta}$, where a and β are the intercept and slope of the demand curve. The formula for the estimated value of WTP is:

$$\hat{WTP} = -q\left(\frac{\hat{\alpha}}{\hat{\beta}}\right) + 0.5q^2\left(\frac{1}{\hat{\beta}}\right). \qquad (1)$$

Testing for the significance of the change in $\hat{WTP}$ for a change in quantity from q_0 to q_1 is done using a t-test. This involves dividing the change in $\hat{WTP}$ by an estimate of the variance of that change:

$$t = \frac{\hat{WTP}_1 - \hat{WTP}_0}{\sqrt{Var(\hat{WTP}_1) + Var(\hat{WTP}_0)}} \qquad (2)$$

The variance of the change in $\hat{WTP}$ is the sum of the estimated variances, assuming that the separate samples are independent (zero covariance).

The exact variance of $\hat{WTP}$ is difficult to calculate since $\hat{WTP}$ is a nonlinear function h of the vector $\hat{\theta}$, where $\hat{\theta} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}$.

$$\hat{WTP} = h(\hat{\theta}) \qquad (1a)$$

Now let $g = \frac{\partial h(\hat{\theta})}{\partial \hat{\theta}}$, the 2x1 vector of derivatives of h with respect to $\hat{\theta}$.

Then an approximate formula for the variance of $\hat{WTP}$ is:

$$Var(h(\hat{\theta})) = g' Var(\hat{\theta}) g \qquad (3)$$

Thus the formula for the estimated variance of $\hat{WTP}$ is:

$$g' Var(\hat{\theta}) g = \left[-\frac{q}{\hat{\beta}} \quad \frac{q}{\hat{\beta}^2}(\hat{\alpha} - 0.5q) \right] \begin{bmatrix} Var(\alpha) & Cov(\alpha,\beta) \\ Cov(\alpha,\beta) & Var(\beta) \end{bmatrix} \begin{bmatrix} -\frac{q}{\hat{\beta}} \\ \frac{q}{\hat{\beta}^2}(\hat{\alpha} - 0.5q) \end{bmatrix}. \qquad (4)$$

Estimates for the variances and covariance of $\hat{\alpha}$ and $\hat{\beta}$ are produced when a least squares regression is run. The vector of derivatives is a combination of the estimated demand parameters and quantity, so that the estimated variance in $\hat{WTP}$ can now be solved.

Simulation of the Two-Way Scope Test

Once a general formulation of the test is in place, it can be used on real demand data to get estimates of the percentage change in quantity required to get significant changes in $\hat{WTP}$. The data used in this test are from a study by Dickie et al. (1987). They compared actual to hypothetical demand for strawberries in Laramie, Wyoming. The actual data consist of the number of pints of strawberries that seventy-two people would be willing to purchase at a variety of prices. We split the sample randomly to produce two independent samples of thirty-six each. From these data,

Table 1 - Ordinary Least Squares Estimates

Sample	1	2
Estimate	Q = 1.449- 0.7619P	Q = I.767-I.OP
t Values	(3.062) (-1 .854)	(3.926) (-2.559)
Mean Quantity	0.6111	0.6667
Var(a)	0.2241	0.2025
Var(b)	0.1689	0.1527
Cov(a,b)	-0.1858	-0.1679
Sample Size	36	36

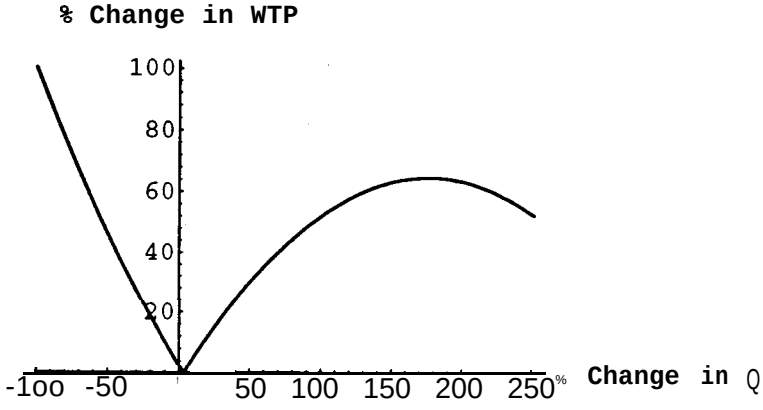
ordinary least squares estimates were produced regressing quantity on price ($Q = \hat{\alpha} + \hat{\beta}P + \hat{\varepsilon}$). These estimates are shown in Table 1.

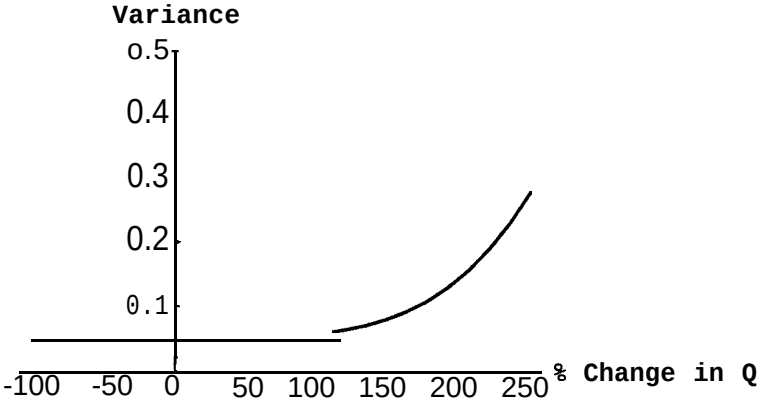
Using these data, the t statistic for changes in WTP caused by changes in quantity in the positive and negative direction was calculated, using Sample 1 to find initial WTP and Sample 2 to find WTP after the change in quantity. The critical t value for the 5% significance level (n = 36) is 2.03, and for the 1% significance level is 2.726. For positive changes in quantity a 103% increase was required to reach the 5% significance level, while the 1% significance level was never reached. For negative changes in quantity, a 46% decrease was necessary to achieve 5% significance, while a 59% change was needed for 1% significance. To ensure the results were not dependent upon the ordering of the samples, this test was also run in the reverse direction (using Sample 2 to calculate initial WTP and Sample 1 to calculate WTP after the change in quantity). Similar results were obtained. [1]

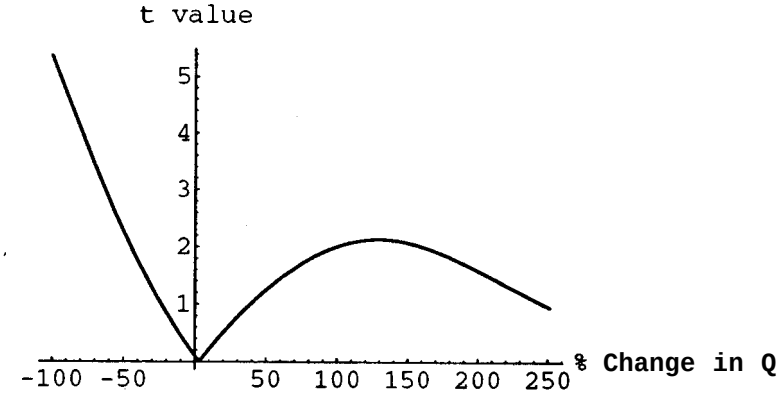
[1] The results in the text are for Test 1. The results of Test 2 are shown here. For positive changes in quantity, significance was not achieved at the 5% or the 1% levels. For changes in the negative direction, a 47% change was required for 5% significance, while a 56% change was sufficient for the 1% significance level.

Figure 2 shows how the % change in WTP, the variance, and the t-value change as the quantity varies from the average consumption level.

Figure 2 - Change in WTP, Variance and t Values for a Private Good







Looking at the bottom panel of Figure 2, it can be seen that the scope test on data for a market good reaches significance more easily with negative changes in quantity than with positive changes. These results are easily explained by referring to the components which make up the t statistic. The change in WTP for a given change in quantity is greater in the negative direction than in the positive direction as is shown in the top panel of Figure 2. An additional explanation for the greater difficulty in achieving significance for an increase in scope is that the variance term in the t statistic increases as q increases as shown in the second panel. Combining this result with the decreasing % change in WTP in the positive direction gives the relation between change in quantity and the t value shown in the bottom panel.

Adjusting the Results for Environmental Goods

The previous section used a private good, strawberries, and ran a two-way scope test on WTP values derived from the estimated demand curve. The next step is to attempt to relate these results to an environmental good. One such environmental good is air pollution for which WTP values can be obtained from the Brookshire, et al. study on visibility in the Los Angeles Basin (1 982). This study asked 186 individuals for their WTP for increasing visibility from fair to good. A hedonic housing market study verified the contingent valuation estimates.

It was shown earlier that the three factors likely to affect significance are variance, slope of the demand curve, and starting point on the demand curve. The two latter factors are unobserved for the environmental data, but one way to compare the likely results of running

the scope test on an environmental good is to make the strawberry data similar to the visibility data with respect to variability.

The data on willingness-to-pay for visibility, the environmental good, result in a mean WTP and a standard deviation for WTP. One problem with a direct comparison between WTP for strawberries, the private good, and WTP for the environmental good is the metric: for the private good we can calculate WTP and the variance of WTP for a known percentage increase (or decrease) in quantity. In the case of the public good, the change is from "fair" to "good."

To circumvent this problem we assume that the standard deviation in WTP is proportional to the mean, that is, that the coefficient of variation (CV) is constant. We then adjust the standard deviation of WTP for the private good so that the CV for the private good is equal to the CV for the environmental good. For the data at hand, this resulted in an adjustment factor of approximately six[2]. That is, the standard deviation of estimated WTP for the private good would have to be multiplied by the adjustment factor so that both the environmental and private good had the same CV. This means that the t statistic must be divided by the adjustment factor. This can be seen by examining equation 4, which estimates the variance of WTP from regression estimates. The quantity $g'Var(\theta)g$ can be written

$$g' \frac{\sigma^2}{n} \left(\frac{X'X}{n}\right)^{-1} g, \qquad (5)$$

[2]The CV for the private good was 0.19, while the CV for the environmental good was 1.13, resulting in an adjustment factor of 5.95 in this case.

where X is the matrix of explanatory variables (simply a constant and price in our example) and s^2 is the variance of the error term. Other things constant, the denominator of the t-statistic is proportional to the square root of s^2, due to the factor s^2/n. Using the adjustment factor on the private good data, no quantity change in the positive direction is large enough for a statistically significant change in WTP.

We now ask the following question: with the standard deviation of an environmental good, how many observations would be necessary to reject the null hypothesis of no difference in WTP for a increase in quantity that satisfies the non-trivial scope test requirements of the proposed regulations? A quantity change of 40% is used, which has been estimated to be an approximate maximum increase in scope allowed under the proposed regulations (the incorporation of other estimates into this analysis is straight forward).[3] Again refer to equation 5. Imagine adding "identical" observations, that is, increasing the sample size in such a way as to leave unaffected both the estimates of a and b (the elements of g) and the second moment matrix, X'X/n. Again the factor s^2/n is important, as all else remains unaffected. This means that the t-statistic is proportional to $\sqrt{n}$. Since the t statistic for a 40% increase in the private good study averaged approximately one for a positive change in quantity, to make the change in WTP significant would require an increase in sample size by a factor of approximately two[4]. In the presence of the six-fold

---

3The estimate is made in McCleland et al.,(1 994). To our knowledge this is the only estimate of the restrictions this test places on allowable quantity changes. The commodity they use is a bid to avoid a hypothetical loss.

[4]For a positive change in quantity of 40%, Test 1 (see Footnote 1) had a t value of 1.03, while the t value for Test 2 was 1.09, for an average of 1.06. For a negative change in quantity of 40%, Test 1 had a t value of -1.74, while the t value for Test 2 was -1.62, for an average of -1.68. Since the positive change requires the larger sample size, it is referred to in the text.

increase in s associated with the environmental good, the sample size must be increased by a factor of approximately $(2 \times 6)^2 = 144$.

In our illustration more than 5,000 observations for each group would be needed (instead of the 36 observations in each group in the market demand based study). Thus, under our assumptions, a minimum sample size of more than 10,000 respondents would be required to demonstrate a significant increase in WTP for an increase in scope of 40% from the average quantity currently consumed.

REFERENCES

Brookshire, David S., Mark A. Thayer, William D. Schulze, and Ralph C. d'Arge, *The American Economic Review,* March, 1982, Volume 72:1.

Dickie Mark, Ann Fisher, and Shelby Gerking, "Market Transactions and Hypothetical Demand Data: A Comparative Study, *Journal of the American Statistical Association,* March 1987, Vol. 82:397.

McClelland et al. "An Examination of Performance Testing Requirements for Contingent Valuation," report to USEPA under Cooperative Agreement #CR-821980, September 1, 1994.

APPENDIX D. COMMENTS ON PROPOSED NOAA SCOPE TEST BY PROFESSORS KENNETH ARROW, EDWARD LEAMER, HOWARD SCHUMAN, AND ROBERT SOLOW

The attached memorandum to NOAA was written by Professor Kenneth Arrow of the Department of Economics at Stanford University, Professor Edward Leamer of the Graduate School of Management at the University of California at Los Angeles, Professor Howard Schuman, Director of the Survey Research Center at the University of Michigan, and Professor Robert Solow of the Department of Economics at the Massachusetts Institute of Technology. All of them were members of the NOAA "Blue Ribbon Panel"; Professors Arrow and Solow are recipients of Nobel Prizes in Economics.

TO: Damage Assessment Regulation Team
c/o NOAA/DAC
SSMC #4
1305 East-West Highway
10th Floor, Station 10218
Silver Spring, MD 20910-3281

RE: NOAA Proposed Rule on Natural Resource Damage Assessments

FROM: Kenneth Arrow, Edward Learner, Howard Schuman, Robert Solow

---

The recently proposed NOAA regulations for contingent valuation surveys includes a "scope test" which is intended to assure the "reliability" of the survey results. (Exhibit 1) This proposed test is apparently a response to the Report of the NOAA Panel on Contingent Valuation, which is excerpted in Exhibit 2. We believe that there is a very sharp conflict between the basic character of the proposed scope test and the sense of the NOAA panel. Because of this difference, we do not think that this test is a proper response to the Panel report, We fear that the proposed test will increase the cost of the surveys with no compensating increase in their "reliability."

The report of the NOAA panel calls for survey results that are "adequately" responsive to the scope of the environmental insult. The proposed scope test is built to assure that there is a statistically detectable sensitivity to scope. This is, in our opinion, an improper interpretation of the word "adequately." Had the panel thought that something as straightforward as statistical measurability were the proper way to define sensitivity, then we would (or should) have opted for language to that effect. A better word than "adequate" would have been "plausible": A survey instrument is judged unreliable if it yields estimates which are implausibly unresponsive to the scope of the insult. This, of course, is a judgment call, and cannot be tested in a context-free manner, as would be the case if the proposed scope test were implemented.

These two definitions will not generally yield the same conclusions. There will be settings in which estimates made with plentiful observations are "statistically" sensitive to the scope but at the same time are "implausible" insensitive. Also, if the sample size is small and the scope difference minor, the estimates may be "statistically" insensitive to the scope, yet "plausibly" sensitive.

The fundamental problem with any purely statistical definition of sensitivity is that it depends (foolishly) on the sample size. In small samples, no effects are "statistically significant." In large samples, everything is "statistically significant." What this means is that the proposed scope test can probably be passed if the trustees are willing to pay a high enough cost. But the willingness to bear this cost has no obvious implications for the "reliability" of the results.

Exhibit 1

Proposed Regulation
Federal Register, Jan. 7, 1994, p.1183.

Scope Test. . . . the trustee(s) shall demonstrate statistically that the aggregate WTP across all respondents for the prevention or restoration program increases (decreases) as the scope of the environmental insult is expanded (contracted) . . . . The demonstration shall be conducted through the use of split samples.

Maximum amount of the difference between scenarios. . . . Prior to the performance of the test, the trustee(s) shall demonstrate that not more than ninety-five percent of respondents in a pre-test or in focus groups indicate that there are meaningful value differences between the scenarios to be tested in any pairwise comparison. The demonstration shall be based on a minimum of sixty valid responses.

Exhibit 2

Report of the NOAA Panel on Contingent Valuation
Federal Register / Vol. 58, No. 10/ Friday, January 15, 1993/proposed Rules

■ Deflection of Transaction Value: The survey should be designed to deflect the general “warm-glow” of giving or the dislike of "big business” away from the specific environmental program that is being evaluated.

■ Burden of proof: . . .If a survey suffered from any of the following maladies, we would judge its findings “unreliable”:

- . . .
- Inadequate responsiveness to the scope of the environmental insult.

APPENDIX E. LETTER FROM DR. DONALD DILLMAN, CURRENTLY CHIEF SURVEY METHODOLOGIST, U.S. CENSUS BUREAU

The attached letter was written by Dr. Donald Dillman, who held at the time and still holds a dual appointment as Chief Survey Methodologist at sthe U.S. Census Bureau and Director of the Social and Economic Sciences Research Center at Washington State University. The views expressed are his own and do not necessarily represent those of the Census Bureau.

Washington State University
Social and Economic Sciences Research Center

Wflson Hall 133
Pullman. WA 99164.4014
509-335-1511
FAX 509-335-0116

March 31, 1993

Mr. Alan Carlin
Office of Policy, Planning and Evaluation
United States Environmental Protection Agency
Washington D.C. 20460

Dear Mr. Carlin:

In your letter of February 12, 1993, you asked by opinion about the recommendation against the use of mail surveys and support of much costlier techniques such as in-person interviews, by the NOAA panel of Nobel laureates in economics and others.

I have read the relevant sections of their report, especially pp.30 and 46-48, and the sections headed "Personal Interview." I'm sympathetic to several of the points the panel raised about the inadequacies of mail surveys, but also believe that they have glossed over, and even ignored some of the difficulties with personal interview and telephone surveys.

One of the panel's objections to mail surveys is the sample frame problem, and this concern is in some cases legitimate. In general there are no readily available household lists for conducting national mail surveys, so that non coverage is a major source of error. The problem is not as bad as they have implied however. They assume the most general case of all adults in the U.S. an urban area, or a state, and then suggest that half the U.S. Population will not be in telephone directories. I don't know from where they obtained that number, but it is higher than ones I have seen, except for southern California. Their assumption of a 75 percent response rate from the remainder is reasonable. They overlook that voter registration lists and drivers license lists are available from many states. It is also the case that contingent evaluation surveys are sometimes done using lists that are quite adequate, e.g. people with hunting or fishing licenses.

Secondly, the report concludes that mail questionnaires will elicit biased answers because of appealing only to those most interested in a natural resource issue, or on one side or the other of the issue. This problem can be dealt with to some degree by obtaining high response rates and through careful questionnaire design. The panel does not recognize that such topical appeal can be a problem with telephone and face-to-face interviews. In fact it is a problem with these methods, especially now that so much more non response to telephone

happens during the course of the interview rather than just being concentrated at the beginning of the interview.

Third the report indicates that it is impossible to guarantee random selection within households or to confine answering to a single respondent, and that it is difficult to control question order effects. This issue is fairly complicated and there is a real lack of data on this concern. It is also impossible to guarantee random selection by the other methods, and when it is close to being achieved it is often off-set in part by lower response rates, because of the very threatening, "How many people live in the household, how many are females (or males), how old are they, etc" sequence that must precede any interviewing. In mail surveys the more common method is to ask for the person with the most recent birthday, and it's unclear how much bias is associated with its use in such surveys. I should also note that for registered voter and other lists, the respondent selection issue they raise is irrelevant.

Whether only one person answers a mail questionnaire is something we really don't know; a definitive study on that topic simply hasn't been done. However, personal interviewing is not immune to that concern. Interviewers are usually trained to avoid such influences, but I've seen instances in which it is impossible to keep a second person from answering the questions addressed towards the interviewee. More typically, the other person sits there and the interviewer never knows the extent to which a respondent takes that other person into account with their answers.

I was rather disappointed that the report didn't raise the issue of social desirability bias, the tendency to offer answers that are normative or that the respondent thinks the interviewer wants to hear. There is considerable evidence that more such bias exists in telephone and face-to-face interviews than in self-administered surveys. In some of the contingent evaluation surveys I have been asked to comment on, it seemed likely that respondents would give socially desirable answers. Also, there is some evidence that interviewed respondents give more extreme answers to telephone and face-to-face interviews, which when combined with social desirability tendencies may result in substantial bias from the use of interview methods. The report should have recognized these potential problems with interview surveys.

The concern about people most interested in a natural resource issue or who are on one side or the other being more likely to respond to mail surveys is an often stated criticism, but a hard one on which to provide data. If one uses all the available procedures for obtaining a high response rate to mail surveys, I question whether that will be much more of a problem than in the

telephone survey, which is now so easy for reluctant respondents to terminate.

The issue about question order effects is a curious one. The existing published literature suggests that order effects are less of a problem in mail surveys than in interview surveys of either type (although I believe this issue to be far from settled). In any event it's curious how one of the desirable qualities of mail surveys gets turned into a negative feature here.

The recommendation that mail surveys be used only if another supplementary method can be employed to cross-validate the results on a random sub-sample of respondents, is a reasonable one, and could be argued for the other methods as well. The social desirability and extremeness biases that may occur in interview surveys, and seem less likely to occur in mail surveys, argue for the cross-validating of interview surveys.

In summary, there are legitimate reasons for being cautious about the use of mail surveys. However, I don't really think the panel's assessment is either balanced or objective. It should also have dealt with the virtual impossibility of guaranteeing high response rates to face-to-face surveys without paying extremely high costs, and it should have dealt with the noncoverage problems of getting into certain areas of the cities, where prudent interviewers will likely refuse to go, or simply can't get in because of gatekeepers (e.g. a condominium complex). It" should also have dealt with the possibilities of social desirability and extremeness biases. Finally, it should also have dealt with the reality of today's, industry standards for face-to-face and telephone interviewing. Frankly, I worry that a report like this will be used to "legitimate" these methods, and then the actual response rates will be quite low because of the limited resources for doing the studies. It's also likely that important contingent evaluation studies simply won't get done, because they will no longer be practical.

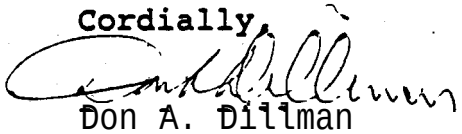
I could imagine a report like this being done at some historical time to argue that a legitimate U.S. Census couldn't be done by mail. (Mail is now relied on for the doing most of the data collection, and has far fewer item non response and perceived measurement problems than does the portion of the census collected by enumerators). I could imagine such a report being used to keep the Current Population Survey which is used to establish unemployment rates, from being done in part by telephone. Had we not learned to use these alternative methods, costs would likely have forced us to do a greatly -abbreviated Census and establish unemployment rates less frequently than monthly, as is now done. Why try to hold contingent evaluation surveys to "standards" to which these two far more important national surveys cannot be held?

It's my conclusion that the report does exhibit considerable bias against mail surveys. Some of the attributed-defects are real, but others are not. More importantly, the report tends to gloss over measurement issues and the problems of producing valid face-to-face surveys results. It assumes the sky-is-the-limit on costs . If the nation needs for contingent valuation surveys to be done, then surely there is a need for making the methodology practical, rather than specifying requirements that will make such surveys only available for those few national problems for which government and large corporations are willing to pay the excessive costs required for the questionable perfection that seems to be demanded by the report.

The direction I would recommend is to think more about mixed mode (more than one method) and "cross-validation" surveys like that mentioned on page 47. Also, I have long sensed that some government and major survey organizations have been-reluctant-to do quality mail survey work, perhaps for reasons similar to why U.S. automobile manufacturers avoided building smaller cars; they are comfortable with face-to-face interviews, and to some extent telephone, because that's what they know how to do. One way of reducing mail survey costs is to build sample frames while doing interviewing for other purposes, but I sense that it's not. something considered very profitable for large survey firms to do, and so far not much effort along these lines has been undertaken.

I hope these comments are helpful.

Cordially,



Don A. Dillman
Professor and Director