Evaluating the Validity of Contingent Valuation Studies

Robert C. Mitchell
Clark University
Worcester, MA

Richard T. Carson
Department of Economics
University of California
San Diego, CA
The contingent valuation (CV) method was developed by resource economists to measure the benefits of difficult-to-value environmental amenities. In contrast to the established methods of travel cost and hedonic price, which use consumer choices in related markets to (indirectly) value nonmarketed goods, CV uses surveys to create hypothetical markets in the good so that respondents can directly express their willingness to pay for changes in the supply of nonmarketed goods under specified contingencies. The CV approach offers two notable advantages over the indirect methods: It directly measures the willingness to pay measure prescribed by welfare theory and it is capable of valuing a much broader range of amenities than the behavior-based methods, including those not yet provided. Providing the results are valid, these advantages make it a potentially attractive methodology for valuing recreation benefits, especially for facilities that are contemplated but not yet provided.

The question of validity— or the degree to which CV estimates are accurate representations of the respondents' true willingness to pay (WTP)—is of vital importance in any assessment of CV as a methodology. This question has been raised in two quite distinct ways. The first is to ask: "Is the method valid?" This is the question asked by many economists who feel much more comfortable with benefit measurement approaches based on what people do than with a survey-based method such as CV that relies on what people say they would do. From these economists' perspective, the latter approach is unlikely to measure people's real preferences because self-interested respondents will be motivated to either give responses shaped by strategic considerations or respond based on a less than optimal effort to search their preferences and arrive at a value for the amenity. The second approach is to ask a more specific question, "To what extent are the findings of a given CV study valid?" This tends to be the approach taken by survey researchers and cognitive psychologists (among others). These researchers are not inclined to dismiss the ability of surveys to measure behavioral intentions out-of-hand, but they know from experience that survey findings, particularly those that involve difficult-to-answer questions, are vulnerable to response effects that occur when the interview process or the wording of the questionnaire influences the respondents' answers.

In this paper, we suggest that the preoccupation of CV researchers in previous years with demonstrating that the method is valid has led to a relative neglect of the important issue of the quality of individual studies. We discuss the consequences of this situation and offer an approach to the assessment of individual CV studies.

THE EFFORT TO VALIDATE THE METHOD

During the past decade, a major concern of CV researchers has been to convince economist skeptics that the hypothetical character of CV surveys is not sufficient grounds for dismissing the method out-of-hand. The result has been that most applications of the CV method, until recently, have had the avowed purpose of assessing its validity. Matters of survey design, sample size, and sampling procedures have taken second place to the implementation of experimental designs using split samples to test for such things as strategic behavior and the effect on the mean WTP amounts of varying such design features as the payment vehicle or the starting point. The results of these experiments have been interpreted as generally favorable to the method.

The most prominent validation strategy, however, has been to show that CV studies do as good a job of measuring benefits as their more favored indirect counterparts. For example, WTP estimates from a CV study have been compared with estimates of an amenity, measured in a series of travel cost and hedonic price studies (Bishop, Heberlein, Welsh, & Baumgartner, 1984; Heberlein & Bishop, 1986) to what people would pay in real dollars for hunting permits in a simulated market. Although suggestive, these comparisons have their limitations. In the first place, it is well recognized that estimates based on the indirect methods, such as a hedonic price estimate of air pollution reduction benefits, are themselves prone to error, owing to the fact that their indirect relationship with the good being valued necessarily requires the use of largely arbitrary assumptions to arrive at a WTP amount (Cummings et al., 1986). Because indirect estimates cannot be regarded as measures of the true WTP amount, they cannot serve as absolute criteria for the validity of a CV measure. Instead, they are parallel measures of the same unmeasured construct. In the language of psychometrics, the greater the correlation between the measures, the greater the evidence for construct (not criterion) validity. This means that the validity of both measures is supported by such a finding, not just the CV measure. A second limitation of the CV-indirect comparisons is that they are restricted to a subset of amenities for which the indirect methods can be used. This includes recreational sites, hunting opportunities, and air visibility in Los Angeles, but excludes national water quality and other public goods for which the CV method is most needed.

These limitations notwithstanding, the credibility ascribed to the indirect measures by resource economists has meant that results of CV-indirect comparisons have assumed the status of a crucial validity test for the CV method. The most extensive assessment of these results was made by Cummings et al. (1986) in their review of 15 hypothetical-indirect comparisons from eight studies. Cummings et al. first developed a "reference accuracy level" based on the accuracy of other benefit measurement methods, which they set at +/−50%. All of the hypothetical-indirect comparisons they examined either met the reference accuracy level or were no higher than 60% of each other. This degree of overall correspondence led Cummings et al. to conclude that the CV method is reasonably accurate, at least when used to value amenities also accessible by indirect methods. In their words:

Assuming that, within the range of +50%, value estimates derived from indirect market methods include "true" valuations by individuals, these results suggest that CVM (contingent valuation method) values may yield "accurate" estimates of value in cases where individuals have had some opportunity to make actual previous choices over that commodity in a market framework. (Cummings et al., 1986, p. 102).

Positive research findings such as these have helped to address the concerns of the skeptics and have played a crucial role in establishing the CV method's credibility, as had other research that established the method's theoretical foundation (Freeman, 1979; Randall, Ives, & Eastman, 1974; Randall &
Stoll, 1980). As a result, several government agencies now certify that contingent valuation is an acceptable procedure for measuring water quality benefits (Water Resources Council, 1979), water project benefits (Moser & Dunning, 1986), and damages under Superfund legislation (U. S. Department of the Interior, 1986). Economists (including journal editors) and policymakers are increasingly inclined to take the findings of CV surveys seriously. In addition, an ever-increasing number of researchers are using the CV method to value nonmarketed goods. We know of more than 100 studies that have used the method between 1965 and 1986 (Mitchell & Carson, in press).

As with any method, studies using the contingent valuation approach vary greatly in quality. In the CV case, this variability is compounded by the newness of the method, the many lacunae in our knowledge of respondent behavior in surveys, the inherent difficulty in obtaining dollar amounts for hypothetical provisions of public goods, the frequent lack of sufficient resources to conduct methodologically adequate CV surveys, and the fact that many people responsible for conducting and reviewing CV studies have had no previous experience or training in conducting sample surveys.

While the early CV researchers’ emphasis on research that would help legitimate the CV method was no doubt essential for the method’s survival, this emphasis on the success or failure of the method as a whole has been at the expense of attention to the methodological quality of individual studies. As noted, the experimental character of the early CV studies and their focus on hypothetical-indirect comparisons did not include attention to such matters as pretesting, sample design and sample size, interviewing procedures, and adjustments for nonresponses. While researchers conducting these early studies often used their data to draw conclusions about aggregate benefits, these estimates were usually treated as suggestive or illustrative rather than definitive.

THE VALIDITY OF INDIVIDUAL CV STUDIES

Now that the CV method has achieved a measure of credibility, there has been a marked shift in the last few years from experimental applications designed to validate the method, to policy applications where the goal is to obtain benefit estimates that can be used in making decisions about the provision of public goods. Policy analysts at the federal and state level are increasingly asking whether the findings of particular CV studies are suitable for use in the regulatory analysis of proposed regulations. Some of these decisions are beginning to be contested in court settings where attorneys and expert witnesses argue the relevance and merits of the CV survey on which the decision is based. As a result, the question “To what extent are the findings of this study valid?” is being asked with increasing frequency.

This is an important question with no simple answers. In this paper, we develop an approach for evaluating individual CV studies on theoretical, policy, and methodological grounds with emphasis on the latter. Before doing so, however, we examine an alternative approach to the assessment of individual studies, which was recently proposed by Cummings et al. (1986).

THE FORMAL CRITERIA APPROACH

Although Cummings et al. (1986) primarily focused on the validity of the CV method, they also recognized the need to assess the quality of individual studies. This led them to propose what might be called the formal criteria approach to this problem. On the supposition that the CV studies used in the CV-indirect method comparisons that met the 50% reference level were acceptably accurate, they advanced the idea that the characteristics these studies share with the “institution underlying indirect market methods” (Cummings et al., 1986, p. 102) may serve as criteria to evaluate the accuracy of other CV studies. They called these criteria “reference operating conditions” (ROCs). They reasoned that if a new CV study meets the full set of ROCs, its findings would be valid in the sense that they could be presumed to be within 50% of the true value.

The ROCs originally proposed by Cummings et al. (1986) were:

1. Subjects must understand, be familiar with, the commodity to be valued.
2. Subjects must have had (or be allowed to obtain) prior valuation and choice experience with respect to consumption levels of the commodity.
3. There must be little uncertainty.
4. Willingness to pay, not willingness to accept, measures are elicited.

These four ROCs were thought to capture the characteristics of market situations, such as buying a new refrigerator or deciding to visit a particular park, which made them amenable to accurate consumer valuation. For example, because consumers normally buy consumer goods rather than sell them, they would be most comfortable with a CV format that asks how much they would pay (ROC 4). ROC 3 is based on research in psychology and experimental economics that shows that when people are uncertain their valuation decisions may be prone to distortions introduced by extraneous aspects of the valuation situation. To the extent that one or more of the ROCs is not satisfied, Cummings et al. (1986) believed the range of accuracy will expand beyond the 50% reference level, reflecting the errors associated with the excluded ROC.

CV researchers and others have debated the validity of the original set of ROCs proposed in Cummings et al.’s (1986) draft report. The ROCs were based on the assumption that a consumer market model is the appropriate model for CV studies. We have argued that in many cases, particularly for public goods, a political market model is preferable because people value public programs and amenities in referenda (Carson, Hanemann, & Mitchell, 1986; Mitchell & Carson, 1986a, in press). Acceptance of a referendum model would imply quite different ROCs; for example, voters often make binding choices about amenities with which they have relatively slight familiarity.

Even if a “correct” set of ROCs were identified, the formal criteria approach to validating individual CV studies would still suffer from fundamental problems. One problem is the usefulness of the 50% accuracy level. On the one hand, this implies a higher level of precision than is warranted. If a genuine criterion was available to compare to a CV estimate, it might be appropri-
ate to accept accuracy levels. However, convergent validity involves comparing two estimates, neither of which can be assumed to represent the unmeasured variable. Although it is reassuring when the two measures of the same concept are close to one another, the possibility exists that both are inaccurate (the confidence economists have in indirect measures notwithstanding). It therefore seems inappropriate to speak confidently about percent levels of accuracy. On the other hand, if one accepts the premise that an accuracy criterion is feasible and desirable, an accuracy range of 50% is probably too large. Applying such an error range to benefit estimates would, in many cases, fail to provide useful guidance to policymakers. Accepting this criteria as the best one can do—the implication of Cummings et al.'s (1986) analysis—is equivalent to feeling satisfied “if one’s bullet hits the target anywhere no matter how far it lands from the bulls eye.” While at times this may be all that is possible, acceptance of a +/- 50% criterion removes the motivation to obtain more accuracy.

The other, and fundamental, problem is that the ROCs dwell solely on formal criteria and omit consideration of various theoretical and methodological factors, which must be taken into account in evaluating the accuracy of individual CV studies. Taken literally, the formal criterion approach seems to imply that as long as the amenity being valued in a CV survey meets the ROCs, any conscientious application of the method will yield acceptable (+/- 50% of the true value) estimates. In fact, a study could meet Cummings et al.’s (1986) original ROCs or any or all of their 11 alternative ROCs and still be inaccurate. To cite an extreme case, estimates based on a haphazard sample of 50 people using a questionnaire that has not been adequately pretested for respondent comprehension, which does not describe the amenity changes properly and is vulnerable to one or more of the many biases to which CV surveys are subject, will not yield useful WTP estimates no matter how familiar the respondents are with the amenity being valued.

### THE METHODOLOGICAL EVALUATION APPROACH

The aim of a CV survey is to obtain a population estimate for people’s WTP for a given amenity. To evaluate a particular study, researchers must scrutinize various factors of the study:

1. The wording of the CV scenario is critical because it provides the stimulus to which the respondents respond. The researcher who designs a CV study creates a hypothetical scenario for the respondent, of which some features, such as the quantity of the good, are intended to be taken into account by the respondent when he or she assesses the value of the amenity. Other features, which may include the government agency that would provide the good or the sequence of the questions, are intended to provide a plausible background for the valuation situation without themselves influencing the valuation outcome.

2. The administration of the instrument is important because the method of administration may introduce errors independent of the scenario’s wording.

### 3. The adequacy of the sample design and its implementation will have an independent effect on the accuracy of the population estimates.

### 4. The sequence in which substitute or complement amenities are valued will influence the size of the WTP amounts. In the following sections, we discuss each factor.

### The Scenario

One of the difficulties in designing a CV instrument is that it must meet the dual criteria of satisfying the requirements imposed by economic theory and the need of respondents for a meaningful and understandable set of questions. Someone who wishes to evaluate a study must have access to the complete text of the survey, as administered. In addition, it is useful to have information about why the researchers designed it to include what they learned from their pretests. Table 1 shows a set of design criteria that must be met by any CV scenario, and the consequences of not meeting them. Each of the five criteria is a necessary but not sufficient condition for a valid scenario; together they may be regarded as necessary and sufficient.

The first two criteria concern the fit between the subject matter of the scenario and the requirements of theory and policy. If, for example, the scenario describes the wrong property right or

| TABLE 1. Scenario Design Criteria and Contingent Valuation Measurement Outcomes |
|---------------------------------|--------------------------------|---------------------|
| Is the Scenario... | If not, respondent will... | Measurement consequence |
| Understandable by respondent as intended? | Value wrong thing (conceptual misspecification). | Measure wrong thing. |
| Plausible to the respondent? | Substitute another condition, or Not take seriously. | Measure wrong thing. Unreliable, bias-susceptible don’t know, or protest zero. |
| Meaningful to the respondent? | Not take seriously. | Unreliable, bias-susceptible don’t know, or protest zero. |
budget constraint, the data are incompatible with economic theory. From a policy perspective, perhaps the most crucial thing is that the scenario adequately describes the amenity change that the policymaker wishes to value. If the findings of a CV study of risk benefits were intended to apply to low level risk reductions, such as from 2 in 1 million to 1 in 1 million, a scenario that describes risks of 1 in 1,000 or even 1 in 100,000 would be misspecified. Similarly, the description of a new recreational area should include all its salient features if the WTP amounts are to properly represent its true value. It is important, in this context, to be aware of the tradeoff between generality and specificity in the descriptions of amenities in CV surveys. The researcher often wishes to apply his or her results to a variety of settings that require findings insensitive to the details of a particular scenario, for example, the location of a recreational area in Ohio rather than Indiana or the use of a utility bill payment vehicle instead of a "higher prices and taxes" vehicle. However, we know that sometimes what seem to be minor changes in the description of an amenity can have large effects on the elicited WTP amounts. Therefore, the closer the fit between the amenity valued in a CV survey and the amenity that a policy analyst wishes to value, the greater the confidence the analyst can have that the CV findings are relevant to the policy decision.

Presuming that the scenario is properly specified from the standpoint of theory and policy, it is necessary to communicate the scenario accurately to the respondents. Conceptual misspecification occurs when respondents understand the scenario in a different way than the researcher intended. This problem tends to be underestimated by researchers untrained in survey research techniques. As Sudman and Bradburn (1982) observed:

The fact that seemingly small changes in wording can cause large differences in responses has been well known to survey practitioners since the early days of surveys. Yet, typically, the formulation of the questionnaire is thought to be the easiest part of the design of surveys—so that, all too often, little effort is expended on it. (p. 1)

For example, some respondents think of "environmental problems" as including trash on city streets and local crime, a broader range of concerns than someone using the term in a survey instrument is likely to want it to mean. Comprehension problems can seriously distort WTP estimates. If some respondents think they are being asked about drinking water in a study that is attempting to ask them about surface water quality in lakes, rivers, and streams, or if they think they are being asked to say what a "fair" price would be for an amenity instead of the highest amount they would pay for it before doing without it, or if they think a risk reduction they are asked to value will reduce the risk to zero when in fact there will still be some risk from a contaminant, the researcher will end up measuring the wrong thing. This places an unusually heavy burden on the designer of a CV survey to undertake a careful and, if necessary, extensive program to try the instrument out under various conditions (pretesting) (Converse & Presser, 1986).

Just because a respondent understands or could understand the scenario does not mean that he or she would necessarily be motivated to make the effort necessary to take the hypothetical situation into account and determine the value of the amenity to him or her. Two factors are particularly important in motivating valid responses to CV scenarios. The first is the plausibility of the scenario. This involves a variety of factors, all of which enhance the perceived realism of the hypothetical market. Is the hypothetical market sufficiently believable to the respondent that he or she will take it seriously? If a good is currently provided for a relatively nominal cost, such as a license to hunt or the use of a state park, respondents may find it difficult to believe that it can have a value significantly higher than these reference amounts even if this is in fact the case. Is it conceivable to the respondent that the outcomes described in the scenario could occur? Some respondents, who do not believe nuclear power can be made "safe," will be incredulous if a scenario asked them how much they would pay for programs to reduce the risk from a given nuclear power plant to nearly zero. Is the choice situation one that makes sense to the respondent? An electric utility bill will be a more plausible payment vehicle than a sales tax for an air visibility scenario because the former has a more understandable connection to the cause of the visibility changes than does the latter. A hypothetical referendum often makes more sense to respondents than a hypothetical private goods market for nonmarketed goods. In all these ways, plausibility reduces the uncertainty in the respondent's mind about the choice situation.

Two undesirable outcomes may occur if the respondent perceives the scenario as implausible. One outcome is that respondents may substitute what they believe to be a more plausible condition for the one described in the scenario. Asked to value a recreational area via a scenario that has the users pay for it, respondents may (consciously or unconsciously) assume that the government would pay for it out of taxes and under value it in their WTP amounts. The result would be a WTP amount for the appropriate good under conditions other than intended by the researcher. The second outcome is that the respondent would not be motivated to take the valuation exercise seriously. To the extent that this occurs, a variety of measurement consequences, none of them desirable and some subversive of accurate benefit estimates, may occur. The respondent might take a "wild guess" at an amount, which would affect the reliability of the WTP estimate, or the respondent might be motivated to minimize the effort involved in answering the valuation question by saying "don't know," by giving a protest zero (i.e., a $0 willingness to pay amount offered to appease the interviewer and which does not represent a true $0 valuation), or by giving a biased WTP amount. A classic example of bias is when respondents' WTP amounts vary systematically according to whether a $1 or $10 amount is used as a starting point for a bidding game elicitation framework.

Bias, in the sense we are using it here, refers to systematic errors. Unlike random error, which is amenable to assessment by sampling and replicating the survey, there is no applicable body of theory by which validity can be assessed (Bradburn, 1982; Carmines & Zeller, 1979) because we lack an explanatory model of the cognitive processes that underlie respondents' verbal self-reports (Bishop, 1981). In these circumstances, the prevention of systematic errors necessarily has an ad hoc character about it, although survey researchers have developed rules-of-thumb, based on experience and a growing body of survey experiments, which serve to minimize bias.

The question of bias is complicated in CV surveys by the general absence of a measurable true WTP value for public
goods, which can be used to assess the validity of a given study. Thus, bias must be inferred from our partial understanding of respondent behavior—if you ask the question this way, people will likely distort their answers—or from evidence in the survey showing that changing the wording of the scenario in ways that are not expected to affect the WTP amounts, in fact does so. The possibility of starting point bias was indicated by theory, suggesting that under conditions of uncertainty, respondents might take initial amounts as information about the “correct” value for the good. The effect was demonstrated in several CV experiments. “Not expected” is a key phrase here, because some differences may in fact be legitimate contingent effects.

This observation requires some explanation, for until recently there was some confusion in the CV literature on this point. It was earlier assumed that only the nature and amount of the amenity being valued should influence the WTP amounts; all other scenario components, such as the payment vehicle and method of provision, should be neutral in effect (Rowe, d’Arge, & Brookshire, 1980). Therefore, according to this view, an experimental finding that the WTP amounts for a given study differ according to whether a utility bill or a sales tax payment vehicle is used was evidence of “information bias.” More recently, Arrow (1986), Kahneman (1986), and Randall (1986) have argued against this view, holding that important conditions of a scenario, such as the payment vehicle, should be expected to affect the WTP amounts. According to their view, which we accept, respondents in a CV study are not valuing levels of provision of an amenity in the abstract, they are valuing a policy that includes the conditions under which it will be provided and the way the public is likely to be asked to pay for it. This notion that a public good does not have a value independent of its method of financing goes back to at least Wickens (1967) and is fully consistent with economic theory.

The uncertainty induced by scenario implausibility promotes susceptibility to bias because respondents in this condition are susceptible to treating supposedly neutral elements of the scenario, such as the starting points, as clues to what the value of the amenity should be. Table 2 summarizes seven types of bias that result from respondents being influenced by the interview situation or treating elements of the contingent market as providing information about the “correct” value for the good. In each case, the respondent’s WTP amount is distorted by the scenario feature in a directional fashion. For example, the undermotivated respondent may assume the amenity is important by the fact that an interviewer has gone to the trouble of asking him or her about it and give a higher amount than he or she would give if properly motivated to express the value it really has to the respondent (importance bias).

Finally, the relevance of the amenity to the respondent also can play a role in motivating thoughtful responses. If the CV study interviews Colorado residents about an expansion in skiing opportunities, it probably will be more difficult to motivate those residents who do not ski to take the study seriously. If so, the same array of measurement consequences described above for implausible scenarios are likely to occur and, because even in Colorado the number of nonskiers is likely to be large, the results could seriously distort the benefit estimates. Interviewer bias, for example, might induce many of these people to say they would be willing to pay a nominal amount to avoid appearing “cheap” in the eyes of the interviewer. Aggregated

<table>
<thead>
<tr>
<th>TABLE 2. Biases From Compliance and Implied Value Cues (from Mitchell &amp; Carson, in press)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
</tr>
<tr>
<td>Compliance bias</td>
</tr>
<tr>
<td>Sponsor bias</td>
</tr>
<tr>
<td>Interviewer bias</td>
</tr>
<tr>
<td>Starting point bias</td>
</tr>
<tr>
<td>Range bias</td>
</tr>
<tr>
<td>Relational bias</td>
</tr>
<tr>
<td>Importance bias</td>
</tr>
</tbody>
</table>

over a large number of nonskiers, annual WTP amounts of 1 or 2 dollars offered by people who really, if they considered the matter, would value the amenity at $0, could substantially bias the estimate upwards.

ADMINISTRATION OF THE INSTRUMENT

Survey instruments may be read to the respondent in person or over the telephone, or they may be sent in the mail with a request to fill out and return. In recent years, the high costs of in-person surveys and methodological developments in telephone survey technology have led the major academic survey research
centers to experiment successfully with telephone interviews, a methodology that commercial polling houses had used for many years (Groves & Kahn, 1979). The sampling problems presented by unlisted telephone numbers have been overcome by the use of computer-based random digit dialing techniques. An even less expensive survey method is the mail survey, which, unlike telephone interviews, permits the use of visual aids. Here, too, methodological advances have improved the technique. It used to be thought that low response rates (e.g., 20-30%) were inevitable in mail surveys, but techniques now available can result in higher (i.e., 50-70%) response rates for mail surveys under some circumstances. However, these techniques require considerably more effort and expense.

Which characteristics of CV questions should influence the choice of method? At least three come to mind. First, CV questions often involve complex scenarios that require careful explanation and that benefit from the use of visual aids and close control over the pace and sequence of the interview. Second, the need to obtain dollar values requires a method that can motivate respondents to make a greater-than-usual amount of effort. Third, the need to extrapolate from the sample to make benefit estimates for populations necessitates the use of survey methods that support techniques to compensate for missing data, a topic we discuss in the next section.

Based on these criteria, the in-person survey, where the interviewer conducts the interview in the respondent’s dwelling, is the method of choice for most CV situations. The physical presence of the interviewer offers the greatest opportunity to motivate the respondent to cooperate fully with a complex or extended interview and the interviewers can be expected to probe unclear responses and provide observational data (Schuman & Kalton, 1985). In-person interviews also lend themselves to the use of various types of visual aids, or “display cards,” to help convey information or complex ideas, and they support techniques to replace missing responses with responses estimated by using missing data techniques.

The large potential cost savings in using telephone and mail surveys has not gone unnoticed by CV researchers. Several have used mail surveys (Bishop & Boyle, 1985; Bishop & Heberlein, 1979; Bishop et al., 1984; Schulze et al., 1983; Walsh, Loomis, & Gillman, 1984), and others have conducted CV surveys by telephone (Carson, Hanemann, & Mitchell, 1986, Mitchell & Carson, 1986b; Oster, 1977; Roberts, Thompson, & Pawlyk, 1985; Sorg & Nelson, 1986; Sorg et al., 1985). Randall, Blomquist, Hoehn, and Stoll (1985) compared all three methods in their study of the national aggregate benefits of air and water pollution control. Not considering cost, what are the tradeoffs between these methods and the more expensive in-person technique?

The greater impersonality of telephone surveys compared with in-person interviews reduces the ability of the interviewer to motivate the respondent. The absence of visual cues during the telephone interview also makes it harder for the interviewer to adjust the interview to the respondent’s circumstances, nor can the interviewer use visual aids to help communicate the scenario. The result is that respondent attention spans for descriptive material are much lower in telephone surveys than in surveys where the interviewer is present and the description can be reinforced by visual aids. This makes it difficult, if not impossible, to maintain respondent interest and attention while communicating even moderately lengthy CV scenarios.

Although mail surveys have an advantage over telephone interviews—the ability to use visual aids—and an advantage over both in-person and telephone interviews in avoiding the possibility of interviewer bias, they suffer from several important shortcomings from the CV point of view. Mail surveys require the respondent to read and understand the description given in the scenario. Unfortunately, the reading level of a surprising number of Americans is quite low. According to the National Assessment of Educational Progress, which conducted a study of literacy among a national sample of 3,600 young adults between the ages of 21 and 25, 6% were unable to read a short sports story in a newspaper, 20% could not read as well as the average eighth grader, 37% could not present the main argument in a newspaper column, and only 43% could use a street map (Kirsch & Jungeblut, 1986). These data underscore reading comprehension problems because the young adult sample has a higher level of education than comparable cohorts of older people. It would appear that unless the scenario in a mail questionnaire is short and simple, or the respondent is reasonably well-educated and also highly motivated, there is an unacceptably large chance that the respondent may miss important details or misinterpret one or more aspects of the scenario. Another set of problems results from the self-administered character of mail surveys. This makes it difficult to use skip patterns, where the choice of follow-up question depends on the respondent’s answer to previous question, or to tailor the interview to the individual respondent’s needs. A well-trained interviewer can pace the interview according to the circumstances of the interview, repeat questions when the respondent indicates puzzlement or uncertainty, and answer respondent questions (within the limits imposed by the interview protocol).

The self-administered mail survey also does not prevent the respondents from browsing through the questionnaire before they start to fill it out. This precludes the use of multiple scenarios where respondents answer each question in a fixed sequence and without knowledge of the following scenarios. Mail surveys also can distort the sample because those who fail to fill out and return the questionnaire are typically those who have the least degree of interest in the amenity being valued.

While in-person interviews are clearly the technique of choice for CV surveys, experience with telephone and mail CV surveys suggest that, except for the sample nonresponse bias problem (which we discuss below), their shortcomings may be largely overcome provided the respondents are familiar with the amenity or the scenario is relatively simple. For example, when Bishop and Heberlein (1979) sent a mail questionnaire to goose hunters, those receiving the questionnaire were well acquainted with the hunting opportunity they were asked about and their nonresponse rate was extremely low for a mail survey. The offshore recreational divers interviewed by Roberts et al. (1985) over the telephone were also familiar with the type of diving amenity they were valuing.

However, as the material becomes more complex and less familiar to the respondents, the results are less satisfactory. We used a relatively simple referendum format in a telephone survey of people’s values for reduced risks of contracting giardiasis from San Francisco’s water supply (Mitchell & Carson, 1986b). In this case, our use of the telephone method involved a clear tradeoff between cost and precision. Despite considerable instru-
ment development by an academic survey research organization experienced in conducting difficult telephone interviews, we found it necessary to omit a number of important aspects of the hypothetical situation from the scenario presentation, which could have been easily incorporated into a personal interview situation.

Irrespective of how a survey is administered, a major requirement is to ensure that the data it obtains are comparable—that is, the information is gathered in a standardized fashion so that one person's answers can be compared with those given by another person. To this end, survey organizations devote considerable care and resources to pretesting questionnaires and training interviewers. Pretesting is the survey equivalent of the test flight. Just as no plane manufacturer would go into production without rigorously testing its latest design, no survey writer would assume that a questionnaire on a new topic, especially if the questionnaire is complex, could be sent directly to the field without careful tryouts under field-like conditions. Even experienced survey practitioners are often surprised when certain questions work better than they had anticipated and others, which they thought were winners, turn out to be fatally ambiguous. Pretests normally consist of an extended period of trial and error experimentation with draft versions of the questionnaire to see which alternative question wordings and orderings work best. If the topic is novel, the pretest process may include preliminary in-depth research, perhaps using focus groups to learn how people conceptualize and talk about the topic (Desvousges, Smith, Brown, & Pate, 1984; Mitchell & Carson, 1986b; Randall et al., 1985).

Comparability also imposes demands on how interviewers conduct themselves in surveys. As Riesman (1958) observed, the basic task of the interviewer is to "adapt the standardized questionnaire to the unstandardized respondents" (p. 267). Except for mail surveys, questioning is a social process. Each interaction between an interviewer and a respondent is unique because of the particular circumstances in which the interview occurs and the personal characteristics of the two participants. To "adapt the questionnaire," without distorting or changing it, the interviewer has to motivate the respondent to enter into a "special" kind of relationship. Sudman and Bradburn (1982) described how interviews differ from ordinary conversations:

The survey interview...is a transaction between two people who are bound by special norms; the interviewer offers no judgement of the respondents' replies and must keep them in strict confidence; respondents have an equivalent obligation to answer each question truthfully and thoughtfully. In ordinary conversation we can ignore inconvenient questions, or give noncommittal or irrelevant answers, or respond by asking our own question. In the survey interview, however, such evasions are more difficult. The well-trained interviewer will repeat the question or probe the ambiguous or irrelevant response to obtain a proper answer to the question as worded. (p. 5)

It is precisely at the point of probing and handling respondent queries that comparability can be lost unless the interviewer rigorously follows instructions not to offer any information or explanations other than those described in the handbook for the study.14

### Sample Design and Implementation

Probability sampling procedures provide surveys with a straightforward way to generalize from the responses of a relatively small number of respondents to much larger populations. These procedures are based on the principle that each economic agency (such as an individual or a household) in the population of interest has a known probability of being selected. Sampling issues had not received much attention in the CV literature until recently, (see Bishop & Boyle, 1985; Desvousges, Smith, & McGivney, 1983; Mills, 1986; Mitchell & Carson, 1984; Moser & Dunning, 1986) despite the fact that they represent a substantial threat to the accuracy of aggregate WTP estimates.

Deciding who to interview for a CV study and how to locate and interview these people involves a series of decisions. First, the researcher must decide how to define the population of economic agents likely to be influenced by the change in the level of the public good. Do they include the residents of a particular town or other geographic area? Are they those who use the amenity? Another choice to be made is whether the agents are to be individuals or households. Next, the researcher must decide how to actually identify or list this population. This list or method of generating such a list is known as a sampling frame. The actual sample is drawn from this list. The third step is to attempt to obtain valid WTP responses from each of the economic agents chosen to be in the sample. Unfortunately, a sizable number of respondents will fail, for some reason, to give valid WTP amounts. These nonresponses can lead to nonresponse and/or sample selection biases unless corrective steps are undertaken. The eventual benefit estimates can become biased as a result of the sampling decisions and procedures at any or all
of these stages. Four types of potential sampling design and execution bias have been identified (Table 3).

Population choice bias occurs when the researcher misidentifies the population whose values the study is intended to obtain. Populations may be defined in terms of the element (e.g., individual recreation), sampling unit (e.g., cars entering recreation areas), extent (e.g., in two counties in northern California), and time (e.g., during July 1988). Choosing the correct population is simplest when the population who will pay for the good (or who would be presumed to pay according to a given payment vehicle, such as a local tax) coincides with the population who will benefit. It becomes more problematic the greater those who pay and those who benefit diverge, for example, in the case of the huge Four Corners Power Plant at Fruitland, New Mexico (Randall et al., 1984). Residents of the area and visitors who come to enjoy the scenery use public good of air visibility without paying the cost of maintaining it. This payment obligation is (would be) borne by those in Los Angeles (and elsewhere) who purchase their electricity from the utility that owns the plant. Nevertheless, area residents and visitors may be a crucial population for a WTP study of the aesthetic benefits of local air visibility, because they directly experience the benefits.

After the population of interest has been identified, the sampling frame must be defined. The frame may be an existing list of the sample units of interest or, more commonly, a method of generating a list. If the population and the sampling frame diverge, sampling frame bias can occur. This type of bias makes it difficult, if not impossible, to accurately generalize the results of the study to the population initially defined by the researcher, even if there are no other problems in carrying out the survey.

The procedures for defining the sampling frame vary according to the type of survey method used—personal, phone, or mail. The sampling frame for in-person surveys of people who live in a given area are normally based on a physical enumeration of geographically-defined, occupied dwelling units. Where the area is large, various types of area stratification and clustering techniques have been developed to reduce the enumeration costs (Cochran, 1977). Nongeographically based populations often pose more difficult problems for in-person surveys. Suppose those who use a beach or visit a park comprise the population of interest. A valid sampling frame should make it possible for the sample to represent the visitors according to the time of day they visit, the day of the week, the season of the year, and, possibly, by how they use the facility. The sampling frame for telephone surveys can either be chosen from the numbers listed in phone books, with the problem of unlisted numbers (both voluntary and involuntary), or, more preferably, the random digit dialing method. This latter method, which selects numbers at random from the universe of usable numbers for the population of interest (see Frey, 1983), ensures that unlisted as well as listed numbers are included in the sample. Mail survey sample frames are based on lists of potential sampling units. They face the problem of obtaining lists of up-to-date addresses for every economic agency in the population of interest. This is often a difficult task for surveys of the general public because of the frequency with which people in our society change their residences. There are likely to be fewer problems of this type when the appropriate sampling frame consists of a current list of addresses held by a government agency, such as the holders of fishing or hunting licenses.

The remaining types of bias—sample nonresponse bias and sample selection bias—occur because of nonresponse. No matter what sampling plan and survey method is used in a CV survey, some level of nonresponse to the WTP questions is virtually inevitable, with the consequence that the number of those who give valid WTP amounts will be smaller than the number of originally chosen sample elements. There are two distinct ways in which a member of the sample can fail to respond to a WTP question. In the first, unit nonresponse (Kalton, 1983), the person (or household) fails to answer the entire questionnaire. This occurs when people cannot be found at home by phone or in-person interviewers, when they refuse to be interviewed when asked to participate by the interviewer, or when those sampled in a mail survey fail to return the questionnaire.

The second way, item nonresponse, occurs when a respondent answers some or most of the questionnaire but fails to answer a particular question of interest, such as the WTP question. With the exception of questions that ask for the respondent's income, item nonresponse rates exceeding 5-7% are relatively rare in ordinary surveys (Craig & McCann, 1978). In CV surveys, however, nonresponse rates of 20-30% for the WTP elicitation questions are not uncommon when (1) the sample is random and therefore includes people of all education and age levels; (2) the scenario is complex; and (3) the object of valuation is an amenity, such as air visibility, which people are not accustomed to valuing in dollars. Up to a certain point, these higher levels of nonresponse to the WTP questions are acceptable or even desirable. It is unrealistic to expect that 95% of a sample will be able and willing to expend the effort necessary to arrive at a well-considered WTP amount for certain types of amenities. Given the choice between having respondents offer unconsidered guesses at an amount or having them say they do not know how much it is worth to them, the latter behavior is preferable, provided appropriate procedures to compensate for the resulting item nonresponse are used.

Unit and item nonresponse results in the loss of valid WTP amounts from those originally chosen for the sample, and both can contribute to sample nonresponse and sample selection bias. If 1,000 households were drawn by probability-based methods for a CV sample and valid WTP amounts were obtained for only 800 of these households, the researcher would have to determine what effect the missing 200 households have on the WTP estimate (i.e., if the values for the 800 people in the realized sample—those for whom valid WTP amounts are available—accurately represent the values for the amenity held by the population from which the original 1,000 household sample was selected). If nonresponse in a CV survey was not associated with the WTP values held by the original sample, the failure to interview some respondents from the original sample would not cause bias (provided the sample size was reasonably large), although it would affect the reliability of the estimates. A lack of association cannot be assumed, however. In the first place, researchers have found that respondent refusal is often associated with lack of interest in the topic of the survey (Stephens & Hall, 1983), and it would seem reasonable that those who are less interested in the amenity stand a good chance of holding different values for it than their more interested counterparts. Second, response rates typically vary across population subgroups, such as lower income people, and there is ample evidence that WTP amounts are often associated with the character-
istics of these subgroups as are other types of survey variables (Kalton, 1983).

To determine whether observed nonresponse results in bias for a given study, two questions need to be asked. One is whether response rates differ across identifiable categories or groups of households (e.g., users vs. nonusers, different educational levels) and the other is whether systematic differences exist between those within a particular group who responded and those who did not. Bias will occur to the extent that these between and within-group differential response rates exist and are related to the value for the good. A given CV study may suffer from a between-group sample nonresponse bias, a within-group sample selection bias, or both. Sample nonresponse bias would occur if, for example, the sample under-represented the proportion of low income households in the population and low income households held different WTP amounts for the amenity than households of other income levels. Even if the proportion of low income households in a study’s sample was representative, the study could still suffer from a sample selection bias—either by differential selection or by a higher rate of item nonresponse—if the low income people who gave usable WTP amounts differed in their preferences for the good from those low income people who did not express their values for the good.

The in-person, telephone, and mail surveys methods have different vulnerabilities to the sample nonresponse and selection biases. In comparison to the other two methods, mail surveys are particularly prone to errors from these sources, especially the latter. This occurs because not only are the unit response rates for mail surveys lower than those for phone or in-person surveys, but also the potential for sample selection bias is higher because of their self-administered character and the concomitant lack of control the researcher has over the process of getting the respondent’s cooperation and eliciting his or her answers.

With telephone and in-person surveys, it is normally possible to assume that the nonresponses are not related to the subject matter of the survey. In the first place, the failure to interview people who are not found at home or who are incompetent in the interview has nothing to do with their personal reaction to the survey’s topic. Second, those who refuse to be interviewed in these types of surveys, usually do so before they know the specific topic of the survey. Third, studies of people who refuse personal or telephone interviews (Smith, 1983; Stinchcombe, Jones, & Sheatsley, 1981) suggest that refusals occur because of general rather than survey-specific reasons.

These assumptions cannot be made for those who receive a mail survey and fail to return it. Unless the recipient throws the package out without opening it, his or her decision whether to respond or not (including the decision to lay it aside for a while) is likely influenced by examination of the cover letter and the questionnaire. Research has shown that the less salient a mail questionnaire is to a potential respondent, the less the likelihood that the respondent will take the time to fill it out and send it back (Heberlein & Baumgartner, 1978; Tull & Hawkins, 1984). In the case of public goods, interest in the subject matter is likely to be correlated with the value the good has to the respondent, thus, nonrespondents to mail surveys may hold lower or even $0 values for the good compared with respondent value of equivalent demographic categories. Thus, mail surveys have a strong potential for sample selection bias—information from those who happen to give valid WTP answers cannot be used to infer or to impute WTP values for the nonrespondents. This is one reason why market research texts (e.g., Tull & Hawkins, 1984) do not recommend their use for general populations.

Some CV researchers have argued that nonresponse bias is not likely to be significant on the basis of the findings of a study conducted by Wellman, Hawk, Roggenbuck, and Buhyoff (1980). This study compared early and late respondents to a CV mail survey about outdoor recreation, which achieved a 70% response rate. The authors argued, on the basis of apparent similarities between these groups on a number of characteristics, that "time, effort, and dollars spent in intensive followups to increase recreation survey response rates might better be expended on other phases of the research process" (p. 171). This finding is an insufficient basis to assume random nonresponse because Wellman et al. did not study the 30% of their sample who failed to respond to the survey. There are no grounds for believing that late respondents to mail surveys such as theirs are a valid surrogate for the nonrespondents and a priori (see above) and empirical (Anderson, Basilevsky, & Hum, 1983) evidence to the contrary.

The Measurement Sequence

CV researchers often want to combine separately measured components of a benefit. These components may be different benefits for separate geographical locations (e.g., the Grand Canyon, the Rocky Mountain Region, and the eastern United States), or benefits for different parts of a larger program (e.g., the air and water quality components of the national environmental program). In a series of important papers, Hoehn, Randall, and Tolley (Hoehn, 1983; Hoehn & Randall, 1982; Randall, Hoehn, & Tolley, 1981; Randall et al., 1985) showed why, under many conditions, independently measured (by contingent valuation) subcomponent benefits cannot be aggregated without overcounting. They also showed that when subcomponent benefits are measured sequentially in the same study, the order in which the subcategories are presented to respondents influences the values ascribed to each, with the goods valued first receiving higher values than later mentioned goods, assuming no other changes in the scenario. The combined entities might be the benefits for a given amenity for different geographical areas, such as air quality in the Grand Canyon, the Rocky Mountain Region, and the eastern United States (geographical sequence aggregation bias), or benefits for different parts of an overall program, such as the air and water pollution control components of the National Environmental Pollution Control Program (multiple public good sequence aggregation bias). They attribute this behavior to the fact that respondents value each good sequentially as if it were a marginal increment to the existing set of environmental amenities that they enjoy instead of valuing it as an initial possible increment.

The essential problem is that a particular policy change is not well specified with regard to another policy change unless the sequence of the two changes is known by respondents. Respondents in an area that has several polluted lakes will value cleaning up the first lake in their area more than the second lake for
several reasons. First, the presence of the cleaned-up, first lake is available as a substitute for a second, cleaned-up lake. Second, the individual’s allocation of money for the first lake reduces the money he or she has available for cleaning up the second lake. If separate CV studies value the lakes individually, however, respondents will treat the lake they are asked to value as if it were the first lake. An overvaluation of the benefits of a combined cleanup will occur if the separate values are added together. If the lakes are valued in sequence in a single study by a sample of respondents, the benefit estimates for the individual lakes (but not the entire set of lakes) will be biased unless the valuation sequence replicates the actual sequence in which the cleanup will occur. It should be clear that any good being valued has a place in a sequence relative to some other goods—either the other good was provided before, at the same time, or later than the good being valued.

Randall et al. (1985) considered the problem of how to disaggregate WTP estimates for all national environmental programs into specific program components such as air and water pollution control. They showed that unless the sequence in which the components will be implemented is known or unless some strong separability conditions on the utility function are met, there is no unique disaggregation of the component values. The essential problem is that the substitution and income elasticities are unknown. Researchers are advised to design CV studies that value a specific good, and to avoid applying values offered in one study to other situations that have significantly different contexts.

CONCLUSION

Contingent valuation surveys are a particularly demanding use of the survey research method because respondents often have difficulty answering the questions. Respondents are likely to be uncertain about the value (in dollars) that they hold for an amenity. The formidable methodological challenge of this undertaking, the lack of expertise most economist practitioners have in survey research techniques, the newness of the method, and the varying degree of resources available to the researcher make it highly likely that the quality of CV-based benefit estimates will vary considerably. The tendency in the CV literature, however, has been to ignore the issue of methodological quality and to treat one CV study as good as another. We argued that this is a legacy of CV researchers’ great emphasis on showing that the method is valid. Now that the method has gained acceptance, and policymakers are beginning to take CV-based benefit estimates into account in making policy decisions, it is vital to pay attention to the validity of individual studies.

Validity refers to the degree to which the benefit estimates from a given CV survey may be regarded as representative of the “true” willingness to pay held by the population to which the estimates are generalized. Owing to the absence of prices for these goods in real or adequate markets, which is why the CV method is needed, validity cannot be established by comparing the survey result against a criterion, nor is it generally possible to establish its convergent validity by comparing the CV estimate with an estimate of the value of the same amenity obtained by an indirect market method, such as hedonic pricing or travel cost. What can and must be done is to assess the study’s methodology with the aim of determining the degree to which its results are free of bias from factors identified in advance as potential sources of error. We offered a framework for undertaking such an assessment, which was organized around four aspects of the research process. To the extent that a study is found to be adequate on these dimensions, the research consumer will have greater confidence in the study’s findings.

We began with a discussion of the scenario used to obtain the WTP amounts and outlined five broad criteria (see Table 1) that a CV scenario must meet if it is to accomplish the researcher’s policy and theoretical intentions. Next, we turned to the process of conducting the survey including the implications of using the in-person, telephone, or mail techniques. We then considered the adequacy of the sample design and its implementation. The fourth aspect was the potential problem posed by the sequence in which amenities are valued when they are substitutes or complements.

Each aspect offers opportunities, often numerous, for error. It is incumbent on researchers to address these issues in the design of the study and in the presentation of its findings. To adequately evaluate a study, the research analyst must have access to the complete questionnaire and reasonably complete information about the sample, for example, the pretesting, the administration of the instrument, the response rate, the techniques used to handle outliers, and the findings of any experimental manipulations incorporated into the study design. Where a particular source of bias is regarded as an especially plausible threat, the researcher is well advised to devise a formal test for its presence. Provided the sample size is adequate, the survey method lends itself to experiments where equivalent subsamples receive different treatments to determine their effect on the WTP amount. CV research has an impressive tradition of conducting such experiments.

The type of assessment we have described will enable the analyst to rule out a large number of possible sources of bias. If validity problems are identified, the study’s estimates can be qualified accordingly. It will, of course, be impossible to rule out error from other unexamined factors, but CV studies are not unique in this regard. Travel cost and hedonic price studies also require careful evaluation and, in the end, an act of trust that their findings are valid. The framework we have presented is based on current knowledge; the more we understand about respondent behavior in CV surveys, the better we will be able to know where the problems in a given study are likely to lie, if indeed, there are problems.

References


Glencoe, IL: Free Press.


**Footnotes**


2. As shown, for example, by the publication of CV studies in the *Quarterly Journal of Economics* (Greenley, Walsh, & Young, 1981), the *American Economic Review* (Brookshire, Thayer, Schulze, & d'Arge, 1982), the *Journal of Political Economy* (Smith & Desvouges, 1987), and the more specialized natural resource economic journals.

3. The Cummings et al. (1986) volume has an unusual structure. It consists of a lengthy draft report written by the editors in which they review the evidence for the method's validity, a series of papers and comments presented by other CV researchers and outside scholars at an all day conference convened to consider their report, and a final "Summary and Conclusions" chapter where Cummings et al. present their final views on the matters covered in the "draft report." What may be termed the "original" ROCs are those set forth in the draft report.

4. To assess the relationship between the ROCs and the accuracy of hypothetical CV markets, Cummings et al. (1986) examined the degree to which each of the seven CV studies involved in the hypothetical-indirect comparisons met the ROCs. The studies varied in terms of the number of CV studies that showed the highest correlation with their indirect counterparts (within 50% or less) satisfied all the ROCs. These analyses involved common and visible things (e.g., air pollution in Los Angeles) and municipal facilities (e.g., schools and community services) in a western boomtown.

5. See Mitchell and Carson (in press) for a further discussion of this issue and a preliminary framework for understanding respondent behavior in CV surveys.

6. This section is based on material in Mitchell and Carson (in press).

7. See Frey (1983) for a discussion of random digit dialing and other aspects of telephone survey methodology.

8. See Dillman (1978, 1983) and Tull and Hawkins (1984) for useful and somewhat different perspectives on mail and telephone survey practice.

9. On the basis of their study, which obtained relatively similar findings for mail and in-person interviews, Randall et al. (1985) concluded that the in-person interviews were not superior to their mail questionnaires. Unfortunately, the response rates they achieved for each methodology were too low (44% for in-person and 36% for mail) to make a definitive judgement on this issue. They also did not address the important sample nonresponse problems which mail surveys are particularly vulnerable to.

10. It may sometimes be possible to mail materials to households before conducting the telephone interviews. See Sorg et al. (1985) for an example.

11. It must be emphasized that standard survey practice forbids interviews from providing ad hoc explanations when respondents look puzzled, or improvising answers to respondent questions. They are instructed to read only the material provided to them, which may, however, include set answers, previously prepared by the researchers, to questions that the pretesting showed might pose difficulties for some respondents. This additional material is only used if the respondent specifically raises the issue.

12. This is why mail and telephone interview techniques are likely to work best for recreational users.

13. Discrete choice formats (where a respondent is offered a single price on a take-it-or-leave-it basis) are usually required under these circumstances with some loss of information and additional complexity in statistical analysis over the continuous choice format.


15. We often use "unit" when "element" is technically the correct term in what follows because of our frequent defining of households as the relevant definition of an economic agent. In this and many other (but not all) instances, the population unit and the population element are equivalent.


17. Approximately 95% to 96% of American households have telephones. Rich (1977) reported that the rate of unlisted numbers in urban areas increased 70% between 1964 and 1977. Groves and Kahn (1979) reported an unlisted rate of 27% for their latest national sample. According to Frey (1983), "when you add new, but unpublished, listings to this figure, it is possible that at any one time, nearly 40 percent of all telephone subscribers could be omitted from the telephone directory" (p. 62).
18 Item nonresponses on WTP questions fall into four general categories: (1) don't knows, (2) refusals, (3) protest zeros, and (4) responses that fail to meet an edit for minimal consistency.

19 Many CV surveys in the literature use relatively small sample sizes (less than 500, often much less). The loss in statistical power may severely limit the ability of such surveys to conduct methodological experiments or to estimate population statistics within a meaningfully narrow confidence interval. These matters were discussed in detail by Mitchell and Carson (in press).

20 The term "nonresponse bias" as used in the survey research literature often refers to both the between- and within-group biases.

21 It should also be clear that the failure to observe a characteristic related to WTP (e.g., income) can change a sample nonresponse bias into a sample selection bias and that obtaining a previously unobserved characteristic can change a sample selection bias into a nonresponse bias. To be more explicit, let:

\[
WTP = f(X, b) + U, (1)
\]

where \(f(X, b)\) is a regression function based on \(X\), a matrix of predictor variables, and \(U\) is a vector of error terms. Sample nonresponse bias occurs when the sample distribution of \(X\)'s differs significantly from the joint population distribution of \(X\)’s, and sample selection bias occurs when the sample distribution of \(U\) differs significantly from the population distribution of \(U\).

22 This presumes, as is the case with many surveys, that the interview topic is described in general terms at the point when the respondents' cooperation is first requested, to avoid this type of bias. For example, the interviewer would say "we are conducting a study of people's views about certain kinds of environmental issues" instead of the more specific "how much people are willing to pay to reduce the risk of cancer from trihalomethane contamination in their drinking water."

23 Undoubtedly some of those who neglect to respond to mail surveys do so for reasons unrelated to the topic. The nature of mail surveys is such, however, that no interviewer is present to record that a potential respondent is sick or has traveled abroad for a month and these nonresponses cannot be distinguished from those who refuse to answer the surveys.

24 For a discussion of the techniques available to compensate for bias due to nonresponse, see Mitchell and Carson (in press).

25 This type of behavior should not be unexpected, because economic theory clearly predicts that one should observe respondents giving decreasing marginal valuations of close substitutes.

26 Where the income effect is small, the second good in the sequence may be valued more than if it had been the first, if the two goods are complements.

27 Walsh, Sanders, and Loomis (1985) conducted a CV study of the benefits of preserving up to 15 wild and scenic rivers in Colorado. They clearly demonstrated the decreasing benefits of preserving additional rivers.

28 Some techniques discussed in Randall et al. (1985) may be used to obtain an upper or lower bound on the WTP amounts independent of sequence if the researcher is prepared to make some fairly weak assumptions about the curvature of the marginal benefit curves.
Amenity Resource Valuation: Integrating Economics with Other Disciplines

Edited by
George L. Peterson
B. L. Driver
Robin Gregory

Venture Publishing, Inc.
State College, PA 16803

1988