Measuring the Demand for Environmental Quality
John B. Braden & Charles D. Kolstad (Editors)

Chapter V

CONSTRUCTED MARKETS

RICHARD T. CARSON

University of California, San Diego

5.1 Introduction[1]

Markets where environmental commodities may be directly bought and sold are scarce. This has led economists to develop techniques such as household production-travel cost analysis (see chapter 3) and hedonic pricing (see chapter 4) in order to infer the value of environmental commodities from transactions for other goods. The alternative approach is to construct markets where environmental amenities may be bought and sold. These markets may be either hypothetical or real. The objective in either type of market is to measure the consumer's willingness to pay or willingness to accept compensation for the environmental amenity of interest.

While hypothetical markets are most often created during the course of a survey interview, the creation of real markets can take several routes. For instance, a city government creates a market for a park when it holds a public referendum to decide whether the community should establish the public park, and a developer creates a market for units with an ocean view when he or she sells otherwise identical units for different prices depending upon whether they do or do not have views.[2] However, most often economists create these markets using groups of test subjects, and for that reason they are sometimes referred to as experimental markets. In this chapter, a term coined by Richard Bishop, "simulated market," will be used to refer to any market in which real money actually exchanges hands for the usually un-

[1] The author wishes to thank W. Michael Hanemann, Kerry M. Martin, Robert Cameron Mitchell, and the editors for their helpful comments. The remaining errors, of course, are those of the author. The author also wishes to acknowledge the financial support of the University of California Water Resources Center, grant W-722, in writing this chapter.

[2] Offering the ocean view as an option with a known price effectively unbundles the ocean view from the structure. The hedonic pricing method is essentially a theoretical and statistical approach to unbundling and pricing a commodities characteristic.

marketed commodity. Perhaps the key characteristics of any constructed market, hypothetical or simulated, is that initially the market is unfamiliar to its participants.

The historic antecedents for using created markets to value commodities date back to at least the 1940s. Ciracy-Wantrup (1947, 1952) advocated the use of survey techniques to determine the demand for environmental commodities, and Bowen (1943) showed how to determine demand for public goods using the results of referenda. The history of test markets in marketing, a close cousin of our simulated markets, is even older. The strongest influences on current work are, however, much more recent. The most well-developed variant of the hypothetical market approach, known as contingent valuation, stems largely from papers by Davis (1963, 1964) and Randall, Ives, and Eastman (1974), while current work on simulated markets derives largely from work in experimental economics by Charles Plott, Vernon Smith, and their associates, and from a paper by Bishop and Heberlein (1979).[3]

Working with constructed markets often makes economists uncomfortable because in doing so they move beyond the usual purview of economics into the realm of other disciplines such as experimental design, marketing, political science, psychology, sociology, and survey research. What has driven economists to use constructed markets is the market's great flexibility, particularly in valuing environmental commodities or aspects of environmental commodities which are difficult, if not impossible, to value using other benefit estimation techniques. In spite of strong attacks by some economists, constructed markets are becoming more and more widely accepted. For instance, contingent valuation, the most frequently used of the constructed market techniques, is endorsed as a benefits estimation technique in the Water Resources Council (1983) guidelines and to a lesser degree by the U.S. Department of Interior (1986) rules for natural resource damage assessment. Contingent valuation is used by a number of federal agencies, such as the Environmental Protection Agency, the Forest Service, the Department of Interior, the National Marine Fisheries Service, and the Army Corp of Engineers; by various state agencies, such as the Alaska Department of Fish and Game, the Colorado Attorney General's Office, and the Metropolitan Water District of Southern California; by major research organizations, such as the Electric Power Research Institute and Resources for the Future; by government agencies in other countries, such as Australia, Canada, and Norway; and by international organizations, such as the World Bank. The number of resource valuation studies based on constructed markets is growing at a rapid rate.

In terms of specific program areas, contingent valuation has been used most extensively to value changes in air quality (e.g., Tolley and Fabian 1988), water quality (e.g., Smith and Desvousges 1986b), and recreation (e.g., Sellar,

[3] See Plott (1982) for a discussion of the history of experimental economics and Mitchell and Carson (1989) for a discussion of the historical development of the hypothetical approaches to valuing nonmarket goods.

Stoll, and Chavas 1985). The technique is also receiving a great deal of attention in the valuation of risk reductions (e.g., Jones-Lee, Hammerton, and Philips 1985). While these are the main application areas to date, a remarkable range of both environmental and nonenvironmental goods have been valued using constructed markets.

Simulated markets for environmental goods have been primarily used to assess the performance of hypothetical markets, with the best examples being the work of Richard Bishop and his colleagues at the University of Wisconsin and that of William Schulze and his colleagues at the University of Colorado. These economists have also focused on comparing the differences between people's willingness to pay (WTP) for welfare changes and their willingness to accept compensation (WTA) measures. Perhaps the largest body of work in experimental economics looks at free-riding behavior (Marwell and Ames 1981; Bohm 1972). Coursey and Schulze (1986) described how the results from laboratory experiments could be used to help develop better contingent valuation methods. Table 5.1 briefly describes a number of representative contingent valuation and simulated market studies.

5.2 Theoretical Foundation

Constructed markets enjoy a very strong theoretical foundation. Depending on the property right assigned, the preferred Hicksian welfare measure can be expressed in terms of either willingness to pay or willingness to accept compensation. Constructed markets, in principle and in contrast to other benefit measurement techniques, can directly obtain WTP or WTA. The other benefit measurement techniques obtain measures of Marshallian consumer surplus that, in many instances, are good approximations of WTP or WTA.[4] Assume, for instance, that an organization or institution is considering an improvement in environmental quality and desires a measurement of WTP (i.e., the Hicksian compensating surplus — see chapter 2). A participant is asked to respond by giving the difference between two expenditure functions:

$$e(\boldsymbol{p},q_0,U_0,\boldsymbol{Q},\boldsymbol{T}) - e(\boldsymbol{p},q_i,U_0,\boldsymbol{Q},\boldsymbol{T}), \tag{5.1}$$

where $\boldsymbol{p}$ is the vector of prices for the marketed goods, q_i is the environmental amenity being changed, U_0 is the initial, or status quo, level of utility to which the respondent is assumed to be entitled, $\boldsymbol{Q}$ is a vector of the other public goods that are assumed not to change, and $\boldsymbol{T}$ is a vector of the participant's taste parameters (Deaton and Muellbauer 1980). The value of the first

[4] Exact measures of WTP or WTA can be obtained using the travel cost or hedonic pricing methods if very strong assumptions can be made about the specification of the utility function (e.g., Hausman 1981). One of the major advantages of using constructed markets is that in many instances it is possible to avoid making specific assumptions about the form of the utility function.

TABLE 5.1
Representative contingent valuation and simulated market studies.

Authors (year)	Good Valued	Research procedure(s)	Elicitation method
Partial list of contingent valuation studies			
Water quality studies			
Carson, Hanemann, and Mitchell (1986)	Water quality bond issue	Telephone	Take-it-or-leave-it
Carson and Mitchell (1988)	National water quality	Personal interview	Payment card
Davis (1980)	Potomac River	Personal interview	Direct question
Gramlich (1977)	Charles River and national water quality	Telephone, personal interview	Take-it-or-leave-it, direct question
Greenley, Walsh, and Young (1981)	Colorado River	Personal interview	Bidding game
Hanemann (1978)	Boston beaches	Personal interview	Bidding game
Loomis (1987)	Mono Lake	Mail	Take-it-or-leave-it, direct question
Oster (1977)	Merrimack River	Telephone	Direct question
Smith and Desvousges (1986b)	Monangahela River	Personal interview	Bidding game, direct question, payment card, contingent ranking
Sutherland and Walsh (1985)	Flathead Lake, Montana	Mail	Direct question
Air quality studies			
Brookshire, Ives, and Schulze (1976)	Siting of plant and visibility	Personal interview	Bidding game
Loehman (1984)	Visibility in San Francisco	Personal interview	Payment card
Loehman and De (1982)	Air pollution control	Mail	Payment card
Rae (1983)	Visibility at national parks	Personal interview	Contingent ranking
Randall, Ives, and Eastman (1974)	Visibility and environmental damage	Personal interview	Bidding game
Ridker (1967)	Air pollution	Personal interview	Direct question
Rowe, d'Arge, and Brookshire (1980)	Visibility in Four Corners Region	Personal interview	Bidding game
Rowe and Chestnut (1989)	Visibility in national parks	Mail	Payment card
Schulze, Brookshire, et al. (1983)	Visibility in Grand Canyon	Personal interview	Bidding game
Tolley and Fabian (1988)	Visibility in Eastern U.S.	Personal interview	Bidding game, direct question

TABLE 5.1 *Continued*

Authors (year)	Good Valued	Research procedure(s)	Elicitation method
Risk studies			
Acton (1973)	Heart attack programs	Mail, personal interview	Direct question
Frankel (1979)	Value of life (airline crash)	Personal interview	Direct question
Hammerton, Jones-Lee, and Abbott (1982)	Statistical life	Personal interview	Direct question
Hammitt (1986)	Food-borne risks	Focus group	Direct question
Jones-Lee (1976)	Value of life	Mail	Direct question
Jones-Lee, Hammerton, and Philips (1985)	Safety	Personal interview	Direct question, bidding game
Mitchell and Carson (1986b)	Trihalomethanes	Personal interview	Direct question
Mulligan (1978)	Nuclear plant accidents	Personal interview	Bidding game
Smith and Desvousges (1986b)	Hazardous waste disposal sites	Personal interview	Direct question
Tolley and Babcock (1986)	Health risks	Mail, personal interview	Bidding game
Land/recreation facilities studies			
Bergstrom, Dillman, and Stoll (1985)	Agricultural land preservation	Mail	Payment card
Bishop and Boyle (1985)	Illinois State Beach	Mail	Take-it-or-leave-it
Daubert and Young (1981)	Instream flows	Personal interview	Bidding game
Majid, Sinden, and Randall (1983)	Public parks	Personal interview	Bidding game
McConnell (1977)	Day at beach	Personal interview	Bidding game
Randall et al. (1978)	Surface coal mine reclamation	Personal interview	Bidding game
Roberts, Thompson, and Pawlyk (1985)	Offshore diving platforms	Mail, personal interview, telephone	Bidding game
Thayer (1981)	Environmental damage	Personal interview	Bidding game
Walsh, Miller, and Gillman (1983)	Ski capacity	Personal interview	Bidding game
Walsh, Loomis, and Gillman (1984)	Wilderness protection	Mail	Direct question

TABLE 5.1 *Continued*

Authors (year)	Good Valued	Research procedure(s)	Elicitation method
Wildlife, hunting, and fishing			
Brookshire, Eubanks, and Randall (1983)	Grizzly bears, bighorn sheep	Mail	Direct question
Brookshire, Randall, and Stoll (1980)	Elk hunting	Personal interview	Bidding game
Cameron and James (1987)	Recreational fishing	Personal interview	Take-it-or-leave-it
Cocheba and Langford (1978)	Waterfowl hunting	Mail	Payment card
Hageman (1985)	Marine mammals	Mail	Payment card
Hammack and Brown (1974)	Migratory waterfowl	Mail	Payment card
Samples, Dixon, and Gower (1986)	Humpback	Focus group	Direct question
Sorg and Nelson (1986)	Elk hunting	Telephone	Bidding game, direct question
Stoll and Johnson (1985)	Whooping crane	Mail, personal interview	Bidding game
Wegge, Hanemann, and Strand (1985)	Recreational fishing	Mail	Take-it-or-leave-it
Partial list of simulated market studies			
Bishop and Heberlein (1980)	Goose permits	Mail	Take-it-or-leave-it
Bishop and Heberlein (1986)	Deer permits	Mail	Take-it-or-leave-it
Bohm (1972)	Free-riding behavior	Laboratory	Direct question
Bohm (1984)	Government	Mail	Direct question
Coursey, Hovis, and Schulze (1987)	WTP vs. WTA	Laboratory	Bidding game
Ferejohn and Noll (1976)	PBS programming	Mail	Iterative ranking of programs
Hoffman and Spitzer (1982)	Coase Theorem	Laboratory	Payoff chart
Knetsch and Sinden (1984)	WTP vs. WTA for lottery tickets	Laboratory	Direct question
Knez and Smith (1989)	WTP vs. WTA for asset units	Laboratory	Direct question
Marwell and Ames (1981)	Free-riding behavior	Mail, telephone	Payoff chart

expenditure function is Y_0, the participant's current income; the value of the second expenditure function is the level of income that solves for U_0 given $\boldsymbol{p}$, q_i, $\boldsymbol{Q}$, and $\boldsymbol{T}$. WTP is defined as the difference between Y_0 and Y_i. Willig (1976) has shown that equation (5.1) can be expressed in an equivalent form known as the *income compensation function*. If WTP is the desired benefit measure, this function, sometimes referred to as the WTP function, is given by

$$\text{WTP}(q_i) = f(\boldsymbol{p}, q_i, q_0, \boldsymbol{Q}, Y_0, \boldsymbol{T}), \tag{5.2}$$

where q_0 is now taken explicitly to be the baseline level of the public good of interest, and the functional form chosen for $e(\cdot)$ or $f(\cdot)$ imposes restrictions on the other. Equation (5.2) forms the basis for estimating a valuation function that depicts the monetary value of a change in economic welfare that occurs for any change in q_i.

Four additional theoretical questions have occupied the attention of contingent valuation researchers. Two of these, the treatment of uncertainty and the decomposition of an agent's benefit from a change in q_i, can be handled in a straightforward manner in a constructed market framework. The other two — should WTP or WTA be used as the measure of economic welfare and how should individual WTP or WTA be aggregated — are not easily resolved because they involve fundamental philosophical issues. Each of these questions is taken up in turn.

Smith (1987b) has shown that uncertainty can be introduced into this framework in a very natural way by replacing the standard expenditure function in equation (5.1) with the concept of a planned expenditure function in order to obtain the desired *ex ante* welfare measure (also see chapter 2). In the simplest sense, the planned expenditure function returns the amount of money just needed *ex ante* to preserve the perceived status quo of expected utility. Because participants in a constructed market naturally take into account both the uncertainty in their demand and any revealed uncertainty of supply when they make their decisions, their responses are consistent with *ex ante* decision making and welfare measures. In contrast, the other benefit estimation techniques must now contend with the need for a technical correction factor known as *option value* (Chavas, Bishop, and Segerson 1986) because they measure *ex post* rather than *ex ante* economic welfare.

Often inspired by the way that various environmental laws are written and by the limitations of the other benefit measurement techniques, researchers who use constructed markets often attempt to disaggregate (or aggregate) WTP/WTA measures obtained from asking the participant to evaluate equation (5.1).[5] The most popular decomposition is between use and existence values. This happens because existence values typically are not measured by other

[5] For example, the Clean Air Act does not allow a monetary value to be placed on health benefits but calls for consideration of economic values for "secondary" benefits such as visibility improvements.

benefit measurement techniques, such as travel cost analysis. The exclusion of existence values creates a bias in the travel cost analysis; the question, of course, is how "big" is the bias. [In chapter 10, Randall uses the expenditure function representation from equation (5.1) to investigate this issue.] Closely related is the issue of how to aggregate or disaggregate benefits over different geographical areas or different policies. This question, too, has an expenditure function representation (see chapter 10 and Hoehn and Randall 1989) and turns crucially on substitution elasticities. One of the key results of the Hoehn and Randall formulation is that it demonstrated the importance of sequence in valuing environmental amenities or disaggregating total value. This is a disturbing finding for policy makers because it means that an environmental amenity does not have a "context independent" value.

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 the participant. Individuals living in an area that has several polluted lakes will place a greater value on the first lake that is cleaned up in their area than on the second. They do this for several reasons. First, each cleaned lake becomes a substitute for subsequent lakes that require cleaning. Second, the individual's allocation of money for the first lake cleaned up reduces the money he or she has available for cleaning up another lake. If separate studies value the lakes individually, however, participants will treat whichever lake they are asked to value as if it is the only lake to be cleaned up. An overvaluation of the benefits of a combined cleanup will occur if the separate values are added up. If the lakes are valued in sequence in a single study, the benefit estimates for the individual lakes — but not for 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 good — either the other good will be provided before, at the same time, or later than the good of interest.[6]

One of the most enduring controversies in constructed markets is whether WTP or WTA should be used as the welfare measure. Many economists thought that this controversy had largely ended with Willig's (1976) results that showed that for a price change, the difference between WTP and WTA was a function of the income elasticity, and that for reasonable values of the income elasticity, the difference between WTP and WTA had to be small. The other benefit measurement techniques, because they were based on estimated Marshallian demand curves, were incapable of directly providing evidence on the difference between the two Hicksian welfare measures. WTP

[6] Although efforts to decompose a WTP response into use value and existence value have probably received too much attention given its policy relevance (because total WTP is already the desired welfare measure) and determining the substitution relationships between environmental amenities has received far too little attention given its large potential policy relevance.

and WTA could, however, be directly measured using constructed markets and the empirical results consistently showed large differences.

These differences helped spawn a great deal of research. Psychologists such as Kahneman and Tversky (1979) put forth theories of why people treated gains and losses asymmetrically, while economists such as Randall and Stoll (1980) extended Willig's work to quantity changes, and Bockstael and McConnell (1980) looked at corner solutions. Bishop and Heberlein (1979) undertook a major experiment to see if the differences were related to the hypothetical nature of contingent valuation, and Coursey, Hovis, and Schulze (1987) looked at how the two measures of value behaved in repeated trials of the simulated market. The number of papers that have attempted to measure both WTP and WTA or rationalize the differences between the two has become quite large.[7]

The most noteworthy recent paper on this topic is Hanemann's forthcoming paper. Hanemann shows that with imposed quantity changes, the theoretical difference between WTP and WTA is governed by the ratio of an income elasticity to a substitution elasticity rather than by an income elasticity alone, as is the case with Willig's price changes. Substitution elasticity refers to the ease with which other market commodities can be substituted for the given public good while maintaining an individual at a constant level of utility. This elasticity of substitution takes a value of zero if no amount of increment in any market goods can substitute for the change in the public good, and a value of infinity if at least one market good is a perfect substitute for the pubic good. It can be shown that the *smaller* the substitution effect (that is, the fewer substitutes available for the public good) and the *larger* the income effect (that is, the greater the income elasticity of demand for the public good) the *greater* the disparity between WTP and WTA. Conversely, if *either* the income effect is zero *or* the substitution effect is infinite, then WTP and WTA must coincide. If the public good in question is unique and the income elasticity of ordinary magnitude, then the difference between WTP and WTA can be quite large. Hanemann's results appear to encompass many of the previous empirical findings. The largest differences between WTP and WTA tend to be observed when the good being valued is unique; repeated "sales" of the good in question, of course, make that good more commonplace.

Hanemann's work is unsettling because it implies that, in contrast to Willig's results, there may be large real differences between WTP and WTA for unique environmental goods. This suggests that the property right chosen is important. While there are some researchers who are hopeful that contingent valuation might one day be able to measure WTA, the current consensus is that WTA cannot now be reliably measured using a contingent valuation survey. The problem in a contingent valuation market is creating either a plausible situation in which the implicit agent who will purchase the good is likely to convey

[7] See Mitchell and Carson (1989) for a review of this literature.

the money to the participant who can sell the good so that the seller's rational response is to set the price so high that the good will not be sold or a situation in which the purchaser has no choice but to purchase the good so that the seller's rational response is to ask for the highest feasible amount and not the minimum WTA.[8]

Sometimes WTP is obviously the correct welfare measure, in which case the task of the designer of a constructed market is simplified. Sometimes, however, WTA appears to be the correct welfare measure. When this is the case, the debate on WTP versus WTA is sometimes decided in favor of WTP based on questionable logic, such as the following: WTA is the correct measure but since it cannot be measured, the researcher should measure WTP instead. This logic was adopted, for instance, in the U.S. Department of Interior (1986) natural resource damage assessment guidelines and was certainly easier to defend before Hanemann's result.

Mitchell and Carson (1989) have argued that perhaps WTP is the correct property rights assignment in many instances where WTA at first appears to be the correct assignment. For instance, the WTA property right may appear to be correct when an electric utility is responsible for an air quality problem in the city where it is located and the people in the city are assumed to have a right to clean air. The WTA question would inquire how much the city's residents would have to be paid to voluntarily accept the poorer quality air. However, the utility is either publicly owned or regulated, so that residents can have better air quality and higher electricity prices or lower electricity prices and poorer air quality. In such an instance, the residents may possess the right to clean air but they have to pay for it through higher electricity prices. Thus, the effective property right is WTP not WTA. Participants in constructed markets appear to have little problem with this concept if they are told how their money will be used to solve the problem. The key property that makes WTP rather than WTA appropriate is that the same group of agents effectively form both sides of the transaction.

Assume that the desired property right specification leads one to choose the ith agent's willingness to pay, WTP_i, as the welfare measure of choice for that agent. Should the aggregate welfare measure used be N, the population size, times the mean WTP or N times the median WTP, M(WTP)? The standard economic welfare, benefit-cost framework (Just, Hueth, and Schmitz 1982) favors N times mean WTP as the measure that is consistent with the potential Pareto improvement criteria. The public choice literature, however, places much more emphasis on a voting criteria in making decisions about public goods. Constructed markets have the good or bad property, depending

[8] Garbacz and Thayer (1983) provide one instance where WTA seems to have been accurately measured. They asked seniors how much they were willing to accept in the form of higher benefit payments in order to voluntarily give up a senior companion program. This paper seems to succeed in measuring WTA because of the credibility of the option of the government maintaining the senior companion program.

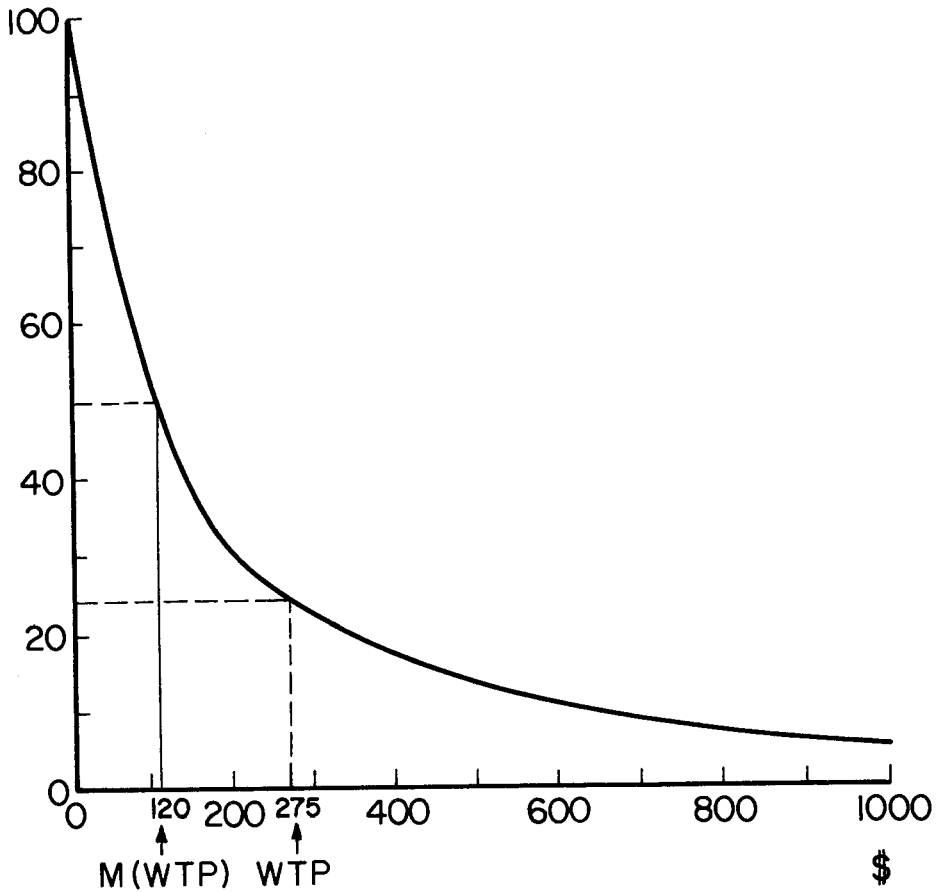

FIGURE 5.1
Percent willingness to pay specified amounts for a fixed quantity of public good.

upon one's perspective, of illuminating the potential divergence between these two criteria because one of the most succinct ways of displaying the results from a constructed market exercise is to display a graph of the distribution of the WTP_i's. Figure 5.1, taken from Carson and Mitchell's (1988) study of WTP for a national clean water program, is typical of the difference between mean WTP and M(WTP) that is often observed. A program that is justified using mean WTP may not be justified using M(WTP). A family of estimators that includes both the mean and the median as special cases is the α-trimmed, where the α largest and smallest observations are given zero weights in calculating the estimate.[9] The statistical properties of the family of the α-trimmed mean estimator are discussed later in this chapter.

[9] It should be emphasized that the observations are not being "thrown way" in calculating

5.3 Designing Constructed Markets

In an ordinary private goods market, a commodity can be bought or sold on a regular basis. Constructed markets have the opposite property. A commodity can only be bought or sold in a constructed market on the terms, including times, defined by whomever set up the constructed market. Constructed markets are of two types: simulated and hypothetical. In a simulated market, the participant makes an "actual" transaction for the good in question. In a hypothetical market, the participant states preferences or makes a pledge about the transaction for the commodity in question. For most purposes, there is no need to distinguish between simulated or hypothetical markets.

Constructed markets may or may not involve experiments, that is, the random assignment of different participants to different treatments, such as different market rules, different market prices, or different commodity characteristics. The term *experimental market* is somewhat of a misnomer as it implies nothing about random assignment of participants to different treatments. This principle of random assignment forms the basis of experimental economics, experimental psychology, and much of statistics. The random assignment of respondents to different treatments within surveys has a long history (Fienberg and Tanur 1985). The topic of experiment design as it relates to constructed markets is taken up in a later section.

A constructed market explicitly or implicitly defines both the payment mechanism and the agent on the other side of the transaction who will deliver or receive the commodity being traded. Three problems are common to the design of all constructed markets: first, structuring the rules of the market in which the good is to be bought or sold; (2) describing the good being valued; and (3) eliciting values or indicators of value in that market. The first two are closely related and are often referred to as the market scenario, which is discussed in this section. The third will be taken up later in this chapter along with other issues, such as market administration, sample design, and estimation of valuation functions.

5.3.1 Market Scenario

How do you tell participants in a constructed market what they are actually buying? Unfortunately, environmental goods such as air quality, water quality, and the risk of toxic chemicals tend to be intangible. In large part, the art of designing constructed markets lies in the description of such goods. The designers of constructed markets have become quite clever in doing this. They

the α-trimmed. The α-trimmed mean is an estimator based on order statistics where the α largest observations are assumed to be offset by the α smallest observations. In doing the trimming, only the rank of the observation is considered and not its absolute value. That is why this estimator becomes more and more resistant to outliers as α is increased.

use photographs to depict different visibility levels due to air quality changes; they denote changes in water quality by what types of water-based recreation are feasible; and they use risk ladders that include familiar activities to inform participants about the effects that changing drinking water standards might have.

Successfully describing the good to be sold is only half of the problem. The other half is to successfully describe a market mechanism under which the good can be sold. The major choice facing the researcher is whether to emulate a private goods market or a public goods market, specifically a referendum situation. The private goods market seems to work well for quasi-public goods, such as duck permits, where exclusion is possible and likely to be desirable. For goods that closely resemble pure public goods, a referendum may be the more logical choice. This choice, however, is not at all neutral. Participants presume the aggregation rule is being used and that other individuals are possibly free riding. Their perception of whether the good can actually be delivered as described is also influenced by the market mechanism used and the description of the agent on the other side of the market.

The wording of the constructed market scenario is critical because it provides the stimulus to which the participants respond. The researcher who designs a constructed market creates a scenario for the participant of which some features, such as the quality of the good, are intended to be taken into account by the participant when he or she assesses the value of the amenity. Other features, which may include the provider of the good or the sequence of questions, are intended to provide a plausible background for the valuation situations without themselves influencing the valuation outcome.

One of the difficulties in designing a constructed market is that it must meet the dual criteria of satisfying the requirements imposed by economic theory and the need of the 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 questionnaire as administered. Table 5.2 shows a set of design criteria that must be met by any constructed market attempting to value an environmental good for policy purposes 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.[10]

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 budget constraint, the data will be incompatible with economic theory. From a policy perspective, perhaps the most crucial aspect is that the scenario adequately describes the amenity change that the policy maker wishes to value. If the findings of a constructed

[10] Even if the scenario is designed correctly, there are other ways in which a constructed market study can fail to obtain valid and reliable data such as from a bad sampling design or faulty execution of the questionnaire.

TABLE 5.2
Scenario design criteria and contingent valuation measurement outcomes.

Is the scenario . . .	*If not, respondent will . . .*	*Measurement consequence*
Theoretically accurate?	Value wrong thing (Theoretical misspecification)	Measure wrong thing
Policy relevant?	Value wrong thing (Policy misspecification)	Measure wrong thing
Understandable by respondent as intended?	Value wrong thing (Conceptual misspecification)	Measure wrong thing
Plausible to the respondent?	Substitute another condition or	Measure wrong thing
	Not take seriously	Unreliable, bias-susceptible DK, or protest zero
Meaningful to respondent?	Not take seriously	Unreliable, bias-susceptible DK, or protest zero

market study of risk benefits was intended to apply to low-level risk reductions, such as from two in one million to one in one million, a scenario which describes risks of one in a thousand or even one in a hundred thousand would be misspecified. Similarly, the description of a new recreational area should include all its salient features if the WTP amounts are to represent its true value. It is important, in this context, to be aware of the trade-off between generality and specificity in the descriptions of amenities in constructed market studies. The researcher often wishes to apply his or her results to a variety of settings that require findings that are insensitive to the details of a particular scenario, such as 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, sometimes what seems to be minor changes in the description of an amenity have large effects on the elicited WTP amounts. Therefore, the closer the fit between the amenity valued in a constructed market study and the amenity a policy analyst wishes to value, the greater the confidence the analyst can have that the findings are relevant to the policy decisions.

Presuming that the scenario is properly specified from the standpoints 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) observe:

> 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.

For example, some respondents think of "environmental problems" as including trash on city streets and local crime. Their definition encompasses a broader range of concerns than was most likely intended by the individual who used the term in the survey instrument. Comprehension problems can seriously distort WTP estimates. The researcher will measure the wrong thing if, for instance, respondents think they are being asked about drinking water in a study that was intended to inquire about surface-water quality in lakes, rivers, and streams; or if they think they are being asked to define a "fair" price for an amenity instead of the highest amount they would pay for it before doing without it; or if they think they are being asked to value a risk reduction that will reduce the risk from a contaminant to zero when, in fact, some risk will remain. This places an unusually heavy burden on the designer of a constructed market study to undertake a careful, and if necessary, extensive program to try out the instrument under various conditions. Converse and Presser (1986) provide one description of this process.

Just because a respondent does understand or can understand the scenario, does not mean that he or she will be sufficiently motivated to take the hypothetical situation into account and determine the value of the amenity to him or her. Two factors, plausibility and relevancy, are particularly important in motivating valid responses to scenarios. *Plausibility* involves a variety of factors, all of which enhance the 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, such as a hunting license or the use of a state park, is currently provided at a relatively nominal cost, respondents may find it difficult to believe that the good can have a value that is significantly higher than these reference amounts even if, in fact, it does. Is it conceivable to the respondent that the outcomes described in the scenario could occur? Respondents who do not believe, for example, that 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 close to zero. Is the choice situation one that makes sense to the respondent? An electric utility bill will be a more plausible payment vehicle than will be 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 does 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.

There are two undesirable outcomes that may occur if the respondent perceives the scenario as implausible. One is that respondents may substitute

what they believe to be a more plausible condition for the one described in the scenario. When asked to value a recreational area via a scenario that has the users paying for it, the respondents may (consciously or unconsciously) assume that the government will pay for it out of taxes, and as a result, undervalue it in their WTP amounts. The result would be a WTP amount for the appropriate good under conditions other than those intended by the researcher. The second outcome is that the respondent will not be motivated to take the valuation exercise seriously. To the extent that this occurs, a variety of measurement consequences may result, none of them desirable and some subversive of accurate benefit estimates. 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 (a $0 willingness-to-pay amount offered to appease the interviewer 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 a $10 amount is used as a starting point for a bidding game elicitation framework.

Bias, in the sense that it is used here, refers to systematic errors. Unlike random error, which is amenable to assessment by sampling and replicating the survey, no applicable body of theory exists by which validity can be assessed (Carmines and Zeller 1979; Bradburn 1982) because there are no explanatory models of the cognitive processes that underlie respondents' verbal self reports (Bishop 1981). In these circumstances, the prevention of systematic error 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.[11]

It is difficult to make a general statement about the likely magnitude of potential biases. The reason is that the threat of various biases is quite specific to the contingent valuation scenario being valued. Most biases in contingent valuation surveys are avoidable; however, some biases, such as starting point bias in a bidding game (which is explained later in the chapter) and sample selection bias in a mail survey, will almost always be present. Typically, most other problems in contingent valuation surveys relate to the people being given inadequate descriptions of what the researchers actually want to value. This can result in large differences between what the researchers actually value and what they intended to value.

The question of bias is complicated in CV surveys by the general absence of a measurable true WTP value for public goods that can be used to assess the validity of a given study. This means that bias must be inferred from the researchers' partial understanding of respondent behavior; for example, researchers know that questions asked in certain ways will likely cause people to distort their answers. Or bias must be inferred from evidence in the survey that shows that changing the wording of the scenario in ways that are not expected to affect the WTP amounts does, in fact, do so. "Not expected" is a key phrase here because some differences may be legitimate contingent effects. The possibility of starting point bias was indicated by theories that suggest 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 experiments.

[11] See Mitchell and Carson (1989) for a further discussion of this issue and a preliminary framework for understanding respondent behavior in CV surveys.

This observation requires some explanation because until recently there was some confusion in the literature on this point. Earlier researchers assumed that only the nature and the 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, and 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, respondents in a CV study are not valuing abstract levels of provision of an amenity; instead, 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 at least to Wicksell's (1967) studies and is fully consistent with economic theory.

The uncertainty induced by implausible scenarios promotes bias because the respondents 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 5.3 summarizes several types of bias that result from the respondents being influenced by the interview 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 directionally by the scenario feature. For example, the undermotivated respondent may assume the amenity is important because an interviewer has gone to the trouble of asking him or her about it. As a result, the respondent will give a higher amount than he or she would if they were properly motivated to express its true value to them (importance bias).

Finally, the *relevance* of the amenity to the respondent can also play a role in motivating thoughtful responses. If the CV study interviews Colorado residents about an expansion in skiing opportunities, it's likely that the interviewers will have more difficulty motivating those residents who do not ski to take the study seriously. If so, the same array of measurement

TABLE 5.3
Typology of potential response effect biases in CV studies.

Incentives to misrepresent responses	
Biases in this class occur when a respondent misrepresents his or her true willingness to pay (WTP).	
Strategic bias	Where a respondent gives a WTP amount that differs from his or her true WTP amount (conditional on the perceived information) in an attempt to influence the provision of the good and/or the respondent's level of payment for the good.
Compliance bias	
Sponsor bias	Where a respondent gives a WTP amount that differs from his or her true WTP amount in an attempt to comply with the presumed expectations of the sponsor (or assumed sponsor).
Interviewer bias	Where a respondent gives a WTP amount that differs from his or her true WTP amount in an attempt to either please or gain status in the eyes of a particular interviewer.
Implies value cues	
These biases occur when elements of the contingent market are treated by respondents as providing information about the "correct" value for the good.	
Starting point bias	Where the elicitation method or payment vehicle directly or indirectly introduces a potential WTP amount that influences the WTP amount given by a respondent. This bias may be accentuated by a tendency to yea-saying.
Range bias	Where the elicitation method presents a range of potential WTP amounts that influences a respondent's WTP amount.
Relational bias	Where the description of the good presents information about its relationship to other public or private commodities that influences a respondent's WTP amount.
Importance bias	Where the act of being interviewed or some feature of the instrument suggests to the respondent that one or more levels of the amenity has value.
Position bias	Where the position or order in which valuation questions for different levels of a good (or different goods) suggest to respondents how those levels should be valued.

consequences described earlier for implausible scenarios are likely to occur, and since 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 in order to avoid appearing "cheap" in the eyes of the interviewer.[12] Aggregated over a large number of nonskiers, annual WTP amounts of one or two dollars, offered by people who really, if they considered the matter, would value the amenity at $0, could substantially bias the estimate upwards.

[12] The best way to avoid interviewer bias, of course, is to get nonthreatening interviewers who have little interest in the actual responses. Graduate students working on the project do not tend to meet these criteria.

TABLE 5.3 *Continued*

Scenario misspecification	
Biases in this category occur when a respondent does not respond to the correct contingent scenario. Except in theoretical misspecification bias, in the outline that follows it is presumed that the intended scenario is correct and that the errors occur because the respondent does not understand the scenario as the researcher intents it to be understood.	
Theoretical misspecification bias	Where the scenario specified by the research is incorrect in terms of economic theory of the major policy elements.
Amenity misspecification bias	Where the perceived good being valued differs from the intended good.
Symbolic	Where a respondent values a symbolic entity instead of the researcher's intended good.
Part-whole	Where a respondent values a larger or a smaller entity than the researcher's intended good.
Geographical part-whole	Where a respondent values a good whose spatial attributes are larger or smaller than the spatial attributes of the researcher's intended good.
Benefit part-whole	Where a respondent includes a broader or a narrower range of benefits in valuing a good than intended by the researcher.
Policy-package part-whole	Where a respondent includes a broader or a narrower policy package than the one intended by the researcher.
Metric	Where a respondent values the amenity on a different (and usually less precise) metric or scale than the one intended by the researcher.
Probability of provision	Where a respondent values a good whose probability of provision differs from that intended by the researcher.
Context misspecification bias	Where the perceived context of the market differs from the intended context.
Payment vehicle	Where the payment vehicle is either misperceived or is itself valued in a way not intended by the researcher.
Property right	Where the property right perceived for the good differs from that intended by the researcher.
Method of provision	Where the intended method of provision is either misperceived or is itself valued in a way not intended by the researcher.
Budget constraint	Where the perceived budget constraint differs from the budget constraint the researcher intended to invoke.
Elicitation question	Where the perceived elicitation question fails to convey a request for a firm commitment to pay the highest amount the respondent will realistically pay before preferring to do without the amenity. (In the discrete-choice framework, the commitment is to pay the specified amount.)
Instrument context	Where the intended context or reference frame conveyed by the preliminary nonscenario material differs from that perceived by the respondent.
Question order	Where a sequence of questions, which should not have an effect, does have an effect on a respondent's WTP amount.

The preceding paragraphs should have clarified that the frequently used term *hypothetical bias* is a misnomer. It's a misnomer because even though the hypothetical nature of the situation may increase the variance of the responses and may make the responses more susceptible to other potentially biasing influences, no evidence exists from WTP studies to suggest a systematic direction for the results of a hypothetical as opposed to a simulated market. Likewise, the frequently used term *information bias* is a misnomer. Participants take into consideration the information available to them in formulating their responses. The problem is that most information likely to be provided to a participant in a constructed market is unlikely to be neutral with respect to willingness to pay for a particular good. In particular, participants have preferences over who provides the good, how it will be provided, and who else will have to pay for it. Therefore, the terms hypothetical bias and information bias should be banished from the vocabulary of constructed market discussions.

5.4 Elicitation Methods

For those who have not actually worked with constructed markets, avoiding strategic behavior and problems with question wording most often appears to be the primary issue in using constructed markets. For practitioners, the central issue is often "how is the valuation response actually going to be elicited." This choice of the elicitation method tends to encompass many of the same issues surrounding threats to reliability and validity.

The most obvious elicitation method is to simply ask someone "What is the most you are willing to pay for this environmental good?" This approach is known as the *direct question method* and it has a number of problems. The major problem is the difficulty that people have answering questions of this type. Difficulty in answering the question tends to manifest itself in one of two ways: a high nonresponse rate and a large number of implausibly high or low answers. Psychologically, people do not usually consider the question "What is my reservation price?" because few real markets operate in this manner. In the typical hypothetical market (and to a lesser degree in a simulated market), a respondent does not have very strong incentives to devote a lot of effort to formulating the correct response to this question, but many people will give an answer, nonetheless. This may result in a larger number of extreme responses, that is, zeros and very large numbers. These problems have spawned the search for a better elicitation method. The direct question method is now most commonly used to value multiple public goods that do not have a natural relationship to each other in terms of WTP.

The second most obvious elicitation method is to start with some WTP amount and in response to "yes" replies, increase that amount progressively

ANNUAL HOUSEHOLD: INCOME BEFORE TAXES
$20,000 - $29,999
(Average annual amount in 1982 taxes and prices paid for some public programs)

$ 0		$190	Police and Fire Protection	$ 620		$1140	
10		210		650		1180	
20		230		680		1220	
30		250		710		1260	
40		270		740		1300	
50	Space Program	290		770		1340	
60		310		800		1380	
70		330		830		1420	
80		350	Roads and Highways	860		1460	
90		380		890		1500	
100		410		920	Public Education	1540	
110		440		950		1580	
120		470		980		1620	
130		500		1010		1660	
140		530		1040		1700	
150		560		1070		1740	Defense Program
170		590		1100		1780	

FIGURE 5.2
Payment card.

until the respondents reply "no." Conversely, one should decrease the amount until a yes response is obtained if the respondent says no to the initial amount. This approach is known as the *bidding game* and was proposed by Davis (1963) and developed to its present form in the classic Randall, Ives, and Eastman (1974) paper. The problem with the bidding game is a phenomena called *starting point bias*. Starting point bias arises from two separate sources. First, the starting point is likely to convey some information about what the value of the good should be, and hence the starting point is likely to influence the magnitude of the respondent's final willingness to pay for the good. The second source, which is the process of getting from the starting point to the respondent's final answer, may influence that answer. If the starting point is far away from the respondent's true value, the respondent may be tempted to prematurely say yes or no to end the bidding, or the respondent may engage in yea saying, or to put it more simply, may agree with the interviewer.

A third method, known as the *payment card* (Mitchell and Carson 1981), gives respondents a card with an array of dollar numbers starting at zero (see figure 5.2). A respondent is asked what number on that card (or a number in between) represents his or her maximum willingness to pay for the good in question. The objective of the payment card is to avoid the awkwardness (that is, high nonresponse rate) of the direct question and the starting point bias problem of the bidding game.[13] The origin of the payment card lies with Hanemann's (1978) checklist and more generally with multiple choice survey

[13] It should be noted that the payment card can subtly introduce its own implied value cue through the range of numbers on the card.

questions. Cameron and Huppert (1987) have raised the issue of whether payment card responses are really people's maximum willingness to pay or whether the amount given by a respondent simply indicates the interval in which his or her maximum willingness to pay lies. Certainly, a checklist or a payment card used in a mail survey has this property and, econometrically, this raises some interesting issues. The appropriate estimator in such a case involves interval censoring and requires one to make some fairly strong assumptions about the distribution of responses within each interval.

The payment card can be used to succinctly inform the respondent about how much they are paying for various other goods. Mitchell and Carson did so in their 1981 study, which first put forth the payment card. Essentially, the choice is one of a classic bias-variance trade-off. Telling respondents what they are paying for some other goods stands a chance of biasing the results. Giving them this information also tends to reduce unexplained variance.

The fourth elicitation method is to obtain a single *discrete response* to a take-it-or-leave-it type of question. In environmental economics, this method stems from the seminal 1979 Bishop and Heberlein paper. Bishop and Heberlein advocated this method because it was easier for respondents to answer and, in particular, easy to implement in a mail survey. To those in the field of public choice, it looked like a referendum.[14] The binary choice format has the advantages of being incentive compatible if two other conditions are met. The first condition is that the participant believes some type of plurality decision rule is being used to make the decision and everyone will have to abide by it. The second condition is that the price is set exogenously and the participant does not perceive his or her answer as influencing the conditions of future choice situations he or she may face.

To implement the simple binary discrete choice approach participants are asked whether they would prefer to have the good at a specified price or do without it. If the participants are individually and randomly assigned to a set of prechosen prices, then it is possible to trace out the percentage of respondents who are willing to pay as a function of price. This approach has two related disadvantages. First, a discrete indicator of the participant's actual willingness to pay is necessary to specify either a utility function, or equivalently, a willingness-to-pay function. Second, a discrete indicator conveys substantially less information than knowing the participant's actual maximum willingness to pay.

There are two major debates over the use of the binary discrete choice elicitation method. The first is over whether one is estimating a random utility

[14] It is necessary, however, to distinguish between a political goods (e.g., referendum) market and a binary discrete choice question because it is always possible to phrase the referendum question in such a manner as to say, What is the most that this referendum could cost you in increased taxes and still have you vote for it?

[15] "Incentive compatible" in this usage means that it is in the participant's selfish interest to say yes if he or she prefers to have the good at the stated price and to say no otherwise. Strategic behavior and truth telling coincide for the rational individual in this case.

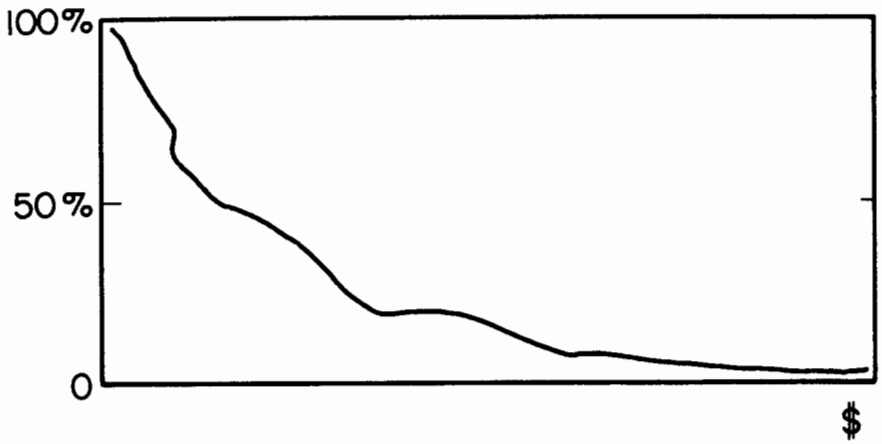

FIGURE 5.3
Percent willingness to pay as a function of required payment.

model (Hanemann 1984b) or a willingness-to-pay function (Cameron and James 1987). The second is over whether it is possible to accurately estimate the mean of the willingness-to-pay distribution from discrete choice data. Both debates revolve around the estimation of the model

$$\Theta(p_i) = f(X,t_i) + \epsilon, \tag{5.3}$$

where p_i is the percentage of respondents willing to pay tax price t_i, X represents respondent characteristics, and Θ is a transformation, possibly linear, of p_i.

The trick, as Bishop and Heberlein (1979) showed, is to estimate the area under the curve defined by equation (5.3) that traces out the percentage of the public that is willing to pay each possible tax price. The vertical axis (figure 5.3) gives the percentage while the horizontal axis depicts the dollars. One of the problems with the discrete choice becomes apparent immediately: the definite integral of the curve defines the mean WTP, but what should the limits of integration be? Setting the lower limit to zero rules out someone having a negative WTP, but most of the time this situation is plausible. Setting the upper limit is more troublesome. In their original study, Bishop and Heberlein set the upper limit equal to $200, the largest dollar amount they asked about in their study.

Let us examine the issue of the upper limit. To make things simple, assume that $\Theta(\cdot)$ is a probit function Φ, $f(X,t_i)$ is linear, and X consists of only a constant term. Equation (5.3) can then be written as

$$\Phi(p_i) = \alpha + \beta t_i + \epsilon. \tag{5.4}$$

In this case, Cameron and James (1987) have shown that WTP = $-\alpha/\beta$. Their approach allows the incorporation of individual characteristics and

Cameron (1988) has extended the approach to cover logit formulations of equation (5.3) as well. Cameron and James seem to avoid the issue of where to truncate the integral, a fact that bothered Bishop and Heberlein. But do Cameron and James really avoid the issue? The answer is no. Cameron and James' major insight is that if t_i is the stimulus variable and t_i is measured in the same unit as WTP, then the estimated coefficient on t_i can be used to recover the scale parameter of the underlying model — a property that is not true in the ordinary probit case. What is less apparent in their paper is that the normal distributional assumption is being heavily exploited in arriving at a closed-form solution for WTP and that this solution implicitly assumes that the upper limit of integration is infinity. Cameron and James have thus provided a very easy-to-use method of estimating WTP if researchers are prepared to make a strong distributional assumption about the shape of the largely unknown tail region. What becomes evident quite quickly in the binary discrete choice models is that the estimate of the median WTP is quite robust to the distributional assumption made and to the transformation of t_i as long as it is restricted to be monotonic.[16]

The other half of the debate revolves around what (5.3) is estimating. Cameron and James see the function $\text{WTP}_i = X_i\beta + u_i$, where X is a vector of respondent characteristics, and they assume that the respondent compares t_i with WTP_i and says yes or no depending on whether WTP_i is greater than t_i or less than t_i. Hanemann (1984b) sees the yes or no response as the result of comparing two indirect utility functions and that estimating (5.3) is justified on the basis of a random utility model. All of this might simply be semantics, but Hanemann shows that the most popular — that is, typically best fitting — form of (5.3), $\Phi(\cdot) = \alpha + \log(t_i) + \epsilon$, is inconsistent with utility theory. This conflict may be resolved in two ways. One is to assume that $\log(t_i)$ is only an approximation to a valid utility function. The other is to assume that every person has a utility function with different parameters and that an equation like (5.4) then, is only a statistical method of describing the population distribution of WTP. Finally, it should be noted that because participants are randomly assigned to a t_i, in large samples t_i will be orthogonal to all individual characteristics so that estimation of the parameter or transformation of t_i is not influenced by the inclusion or exclusion of the participant's characteristics from the estimated equation.

The next issue to be examined is the amount of efficiency that is lost when a discrete choice estimator is used. Alberini and Carson (1990) have recently addressed this issue. They showed that for the simple model given in (5.4), the maximum (Pitman) asymptotic efficiency relative to the discrete choice estimator for the mean WTP relative to any technique that yields observations on actual willingness to pay is approximately $2/\pi$, a little over 60 percent.

[16] Monoticity is probably the weakest restriction imposed by economic theory if equivalent subsamples of participants are assigned to each t_i. All this says is that people prefer low prices to higher prices for the same good.

This means researchers will need at least 66 percent more observations with the simple discrete choice estimator.

This maximum relative efficiency is achieved using Finney's (1971) method of picking the t_i's to minimize the *fudicial confidence interval* — an approximation to a standard confidence interval — around the particular point of interest when the mean and variance of the underlying process are assumed to be known exactly, a priori; in this case, the mean which is estimated by $-\alpha\beta$. Finney's method is fairly robust to a bad guess about the variance; however, relative efficiency falls off dramatically as the guess about the mean deviates from the actual population mean. The other drawback of the Finney approach is that it is highly optimized for estimating a single quantile in the distribution and can do poorly for estimating other quantiles far from the design emphasis.

A second method for determining the location of the t_i's is based on the criteria of D-Optimality (Silvey 1980). The *D-Optimality criteria* is based on picking the t_i's to maximize the Fisher information matrix with respect to the parameters, α and β. The D-Optimality approach has two advantages and two drawbacks relative to the Finney approach. It is fairly robust to bad guesses about the mean but not the variance. It estimates the mean much less efficiently than does Finney's method, but on the other hand, it does much better for estimating quantiles far from the mean.

A third method for choosing the t_i's, given initial guesses for the mean and the variance, is to place the t's at *equal distant quantiles.* The researcher determines how many equivalent subsamples will make up the sample as a whole and assigns a different t_i to each subsample. This method has properties that fall between that of Finney's and the D-Optimal methods and is perhaps the one most natural to standard survey administration procedures.

It is important to note that more subsamples, or equivalently more t_i's, is not preferable to fewer. Finney's method and D-Optimality methods will never yield more than three distinct t_i's. The smaller the subsample, the less precisely estimated is the percentage who will pay the subsample's t_i's. The gain is that the more the t_i's are spread out, the less the risk of a bad guess on the mean. The typical two-point, D-optimal design, under the assumption of normality, places one t_i at approximately $m - 1.14s$ and the other at approximately $m + 1.14s$ (where m is the estimate of the mean of willingness to pay and s is the estimate of the variance of the WTP distribution). Finney's method places them at $m + 0.37s$. With Finney's method, a bad guess on the mean can easily place all of the observations on one side of the t_i's; whereas with D-optimal design, a bad guess on the variance can easily place all of the observations in the center of the two t_i's. For these reasons, the equal distant quantile design seems to be a good compromise for contingent valuation studies. However, even with this latter method, bad guesses for the mean and variance can still dramatically reduce the asymptotic relative efficiency of the discrete choice method to close to zero. This should emphasize the strong

need for pretests to ensure good estimates for the mean and variance. These pretests, at least the initial ones, should probably use an open-ended response format.

Recognition of the inefficiency of the single binary discrete choice question has led researchers to other discrete choice formats. The first of these is best represented by the Bergstrom, Rubinfeld, and Shapiro (1982) paper. They essentially asked respondents a "more, less, or about right" question. And the respondents appeared to be quite able to answer this question. The drawback of the approach is that the statistical model is fairly complex to estimate, and much more specific assumptions have to be made about the form of the utility function.

More in keeping with the simple binary discrete choice question is to repeat it once. Carson, Hanemann, and Mitchell (1986) showed that a Neyman double-sampling scheme could be used to achieve a very large increase in the efficiency of the estimate. If a respondent answered yes to a question, he or she was randomly assigned a higher number and asked again; if a respondent answered no, he or she was randomly assigned a lower number. If repeated often enough, this scheme turns into the bidding game, and thus the source of the inefficiency of the single discrete choice response is made clearer. The single repeat, with a random assignment exploiting the previously revealed preference, seeks to exploit the gain of the bidding game without setting up the yea-saying syndrome or losing the incentive compatibility property.

Seeing that the trick was to narrow the interval where the participant's maximum willingness to pay lay, Carson (1988) and Carson and Steinberg (1989) showed that the appropriate statistical technique was *interval data survival analysis.*[17] Here, price rather than time is the stimulus variable. The variance of the estimates can be shown to be closely related to the width of the intervals and survival analysis easily handles intervals with zero as the left endpoint and right censored endpoints, thus naturally resolving the infinite willingness-to-pay situation that had bothered Bishop and Heberlein. The estimated survival function is simply the estimated demand curve, and the estimated hazard function is closely related to the elasticity of demand. Survival analysis is a well-developed statistical technique. There are survival distributions that force a constant elasticity, such as the exponential; others allow increasing, decreasing, or constant elasticities with respect to price while maintaining monotonicity, such as the Weibull; and still others make it possible to go the complete nonparametric route forcing no restrictions on the shape of the demand curve. Survival analysis can handle covariates and very complicated assignment schemes.[18]

[17] The binary logit and probit models can be shown to be the simplest type of survival model.

[18] Carson (1988) showed that utility theory can be further exploited in double-sampling schemes with certain survival analysis estimators if different amenities asked about have known preference relationships.

5.5 Market Administration

Market instruments may be read to the participants in person or over the telephone, or they may be sent in the mail with a request to complete and return.[19] 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 which commercial polling houses have used for many years (Groves and Kahn 1979). The sampling problems presented by unlisted telephone numbers have been overcome by the use of computer-based random digit dialing techniques.[20] 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 was once thought that low response rates of 20 to 30 percent were inevitable in mail surveys, but techniques are now available that can result in more respectable 50 to 70 percent response rates. These techniques, it should be noted, require considerably more effort and expense.

Which characteristics of constructed market questions should influence the choice of method? First, constructed markets 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 motivates respondents to exert a greater-than-usual effort. Third, the need to extrapolate data from the sample to estimate benefits for various populations requires that researchers use survey methods that support techniques that compensate for missing data — a topic to be considered in the next section.

For most situations, the method that meets all of these criteria is the in-person survey conducted in the respondent's dwelling place. For example, 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 interviewer has the opportunity to probe unclear responses and to provide observational data (Schuman and Kalton 1985). In-person interviews also lend themselves to the use of various types of visual aids, or "display cards," which help to convey complex ideas or bodies of information. Furthermore, they support missing data techniques.

The large potential cost savings in using telephone and mail surveys has not gone unnoticed by constructed market researchers, however. Several have used mail surveys (Bishop and Heberlein 1979; Schulze, Brookshire, et al.

[19] The discussion here refers primarily to contingent valuation surveys. Simulated markets may also be implemented in person, over the telephone, or through mail surveys. Simulated markets sometimes use a variant of personal interviews where individuals are invited into the researcher's lab and a variant of mail surveys. For experimental purposes, a variant of the mail survey is used where students are asked to fill out an in-class questionnaire.

[20] See Frey (1983) and Dillman (1978, 1983) for a discussion of random digit dialing and other aspects of telephone survey methodology.

1983; Walsh, Loomis, and Gillman 1984; Bishop, Heberlein, Welsh, and Baumgartner 1984; Bishop and Boyle 1985) and others have conducted surveys by telephone (Oster 1977; Roberts, Thompson, and Pawlyk 1985; Carson, Hanemann, and Mitchell 1986; Sorg et al. 1985; Mitchell and Carson 1986b; Sorg and Nelson 1986). Randall et al. (1985) compared all three methods in their study of the national aggregate benefits of air and water pollution control.[21] Excluding costs, what are the trade-offs between these methods and the more expensive in-person technique?

First, the more impersonal nature of the telephone survey compared with the in-person interview reduces the ability of the interviewer to motivate the respondent. Second, the absence of visual cues during the telephone interview makes it more difficult for the interviewer to adjust the interview to the respondent's circumstances. In addition, the interviewer cannot use visual aids to help communicate the scenario. The result is that respondents' attention spans for descriptive material are much lower in telephone surveys than in surveys where the interviewer is present. This makes it difficult, if not impossible, to maintain respondent interest and attention while communicating even moderately lengthy constructed market scenarios. It may sometimes be possible to mail materials to households before conducting the telephone interviews. Sorg et al. (1985) provide an example of this.

Although mail surveys have the advantage over telephone interviews of being able 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 when applied to constructed markets. One shortcoming is they 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 percent were unable to read a short sports story in a newspaper, 20 percent could not read as well as the average eighth-grade student, 37 percent could not present the main argument in a newspaper column, and only 43 percent could use a street map (Kirsch and Jungeblut 1986). These data understate reading comprehension problems because the young adult sample has a higher level of education than that of comparable cohorts of older people. Unless the scenario in a mail questionnaire is very 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

[21] 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 percent for in-person and 36 percent for mail) to make a definitive judgment on this issue. Nor did they address the important sample nonresponse problem to which mail surveys are particularly vulnerable.

one or more aspects of the scenario. Another set of problems results from the self-administered character of mail surveys. This causes difficulties in using skip patterns, where the choice of follow-up questions depends on the respondent's answer to previous questions, or in tailoring the interview to the individual respondent's needs. A well-trained interviewer can pace the interview according to the circumstances of the interview and can (within the limits imposed by the interview protocol) answer respondent's questions.[22]

The self-administered character of mail surveys provides no way of keeping the respondents from browsing through the questionnaire before they start to fill it out. This precludes the use of multiple scenarios where it is desired to have the respondents answer the questions in a fixed sequence without knowledge of the following scenarios. Mail surveys can also 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 constructed markets, experience with telephone and mail surveys suggest, except for the sample nonresponse bias problem that is discussed later, their shortcomings may be largely overcome provided the respondents are very familiar with the amenity[23] or the scenario is relatively simple.[24] 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 the nonresponse rate was extremely low for a mail survey. The off-shore recreational divers interviewed by Roberts, Thompson, and Pawlyk (1985) over the telephone were also familiar with the type of diving amenity they were valuing, and consequently, were willing to answer the questions.

As the material becomes more complex and less familiar to the respondents, however, the results are less satisfactory. Mitchell and Carson (1986b) 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. In this case, the use of the telephone method involved a clear trade-off between cost and precision. Even though the survey was developed by an academic survey research organization experienced in conducting difficult telephone interviews, during the interview the researchers had to omit from the scenario

[22] It must be emphasized that standard survey practice forbids interviewers 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 researcher, to questions which the pretesting showed might pose difficulties for some respondents. This additional material is only used if the respondent specifically raises the issue.

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

[24] 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.

a number of important aspects of the hypothetical situation, aspects which could have been easily incorporated into a personal interview.

Irrespective of how it is administered, a major requirement of a survey 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 answer can be compared with the answer given by another. 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, so too, no survey writer would assume that a questionnaire on a new topic, especially if the questionnaire were complex, could be sent directly into the field without careful tryouts under field-like conditions. Even experienced survey practitioners are often surprised when certain questions obtain better results than they had anticipated while others that they thought were winners turn out to be fatally ambiguous. Pretests normally consist of an extended period of trial and error with draft versions of the questionnaire. If the topic is novel, the pretest process may include preliminary in-depth research, perhaps using focus groups (Desvousges, Smith, Brown, Pate 1984; Randall et al. 1985; Mitchell and Carson 1986b; Krueger 1988) to learn how people conceptualize and talk about the topic.

Comparability also imposes demands on how interviewers conduct themselves in surveys. As David Riesman (1958) once observed, the basic task of the interviewer is to "adapt the standardized questionnaire to the unstandardized respondents." Except for mail surveys, questioning is a social process. Each interaction between an interviewer and a respondent is unique owing to the particular circumstances in which the interview occurs and the personal characteristics of the two participants. In order to "adapt the questionnaire" without distorting or changing it, the interviewer must motivate the respondent to enter into a special kind of relationship. Sudman and Bradburn (1982) describe 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 judgment of the respondent's 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.

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.[25]

5.6 Sample Design

Probability sampling procedures provide survey researchers 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 agent, 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 constructed market literature until recently, even though they represent a substantial threat to the accuracy of aggregate WTP estimates.[26] Deciding who to interview for a constructed market 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 who are 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 areas? And does this group include those who use the amenity? Among the other choices the researcher makes 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.* It is from this list that the actual sample is drawn. The third step is to attempt to obtain valid WTP responses from each of the economic agents chosen to be in the sample frame. Unfortunately, there will be a sizable number of respondents who fail, for some reason, to give valid WTP amounts. These nonresponses can lead to nonresponse and sample selection biases unless corrective steps are taken. 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 can be identified. They are summarized in table 5.4.

Population choice bias occurs when the researcher misidentifies the population whose values the study intended to obtain. Populations may be defined in terms of the element, sampling unit,[27] extent, and time. For example, the element could be an individual recreator; the sampling unit, the number of cars entering recreation areas; the extent, two counties in northern California;

[25] The Research Triangle Institute's 1979 publication *Field Interviewers General Manual* offers an informative overview of the interviewer's role and training.

[26] See Desvousges, Smith, and McGivney (1983), Mitchell and Carson (1989), Bishop and Boyle (1985), Moser and Dunning (1986), Edwards and Anderson (1987).

[27] "Unit" is often used although "element" is technically the correct term in what follows because households were frequently defined as the relevant definition of an economic agent. In this and many other instances, the population unit and the population element will be equivalent.

TABLE 5.4
Potential sampling and inference biases in CV surveys.

Sample design and execution biases	
Population choice bias	Where the population chosen does not adequately correspond to the population to whom the benefits and/or costs the provision of the public good will accrue.
Sampling frame bias	Where the sampling frame used does not give every member of the population chosen a known and positive probability of being included in the sample.
Sample nonresponse bias	Where the sample statistics calculated by using those elements from which a valid WTP response was obtained differ significantly from the population parameters on any observed characteristic related to willingness to pay; this may be due to unit or item nonresponse.
Sample selection bias	Where the probability of obtaining a valid WTP response from a sample element having a particular set of observed characteristics is related to their value for the good.
Inference biases	
Temporal selection bias	Where preferences elicited in a survey taken at an earlier time do not accurately represent preferences for the current time.
Sequence aggregation bias	
Geographical sequence aggregation bias	Where the WTP amounts for geographically separate amenities that are substitutes or complements are added together to value a policy package containing those amenities, despite the fact that the amenities were valued in an order (for example, independently) different from the appropriate sequence.
Multiple public goods sequence aggregation bias	Where the WTP amounts for public goods that are substitutes or complements are added together to value a policy package containing those amenities, despite the fact that the amenities were valued in an order (for example, independently) different from the appropriate sequence.

and the time, July 1988. Choosing the correct population is simplest when the population who will pay for the good, or who is presumed to pay according to a given payment vehicle such as a local tax, coincides with the population who will benefit. The greater the divergence between those who pay and those who benefit, the more problematic it becomes to choose the correct population. Consider the case of the huge Four Corners power plant at Fruitland, New Mexico, (Randall et al. 1974). Residents of the area and visitors who come to enjoy the scenery use the 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 the crucial population for a WTP study of the aesthetic benefits of local air visibility because they experience the benefits directly.

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 conducting the survey.

The procedures for defining the sampling frame vary according to the type of survey method used — personal, phone, or mail.[28] 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 dwellings. When the area is large, various types of area stratification and clustering techniques have been developed that make the enumeration costs manageable (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 very real problem of unlisted numbers (both voluntary and involuntary),[29] or more preferably, from random digit dialing. The latter method, which selects numbers at random from the universe of usable numbers for the population of interest (Frey 1983), ensures that unlisted as well as listed numbers are included in the sample. Mail surveys' sample frames are based on lists of potential sampling units. With this method, researchers face the problem of obtaining lists of up-to-date addresses for every economic agent in the population of interest. This is often difficult for surveys of the general public because people in our society frequently change their residence.[30]

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

[28] For nontechnical descriptions of sampling frame development procedures see Sudman (1976) or Tull and Hawkins (1984).

[29] Approximately 95 to 96 percent of American households have telephones. Rich (1977) reports that the rate of unlisted numbers in urban areas soared 70 percent between 1964 and 1977. Groves and Kahn (1979) report an unlisted rate of 27 percent 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."

[30] There are likely to be fewer problems of this type where the appropriate sample frame consists of a current list of addresses held by a government agency as the holders of fishing or hunting licenses.

entire questionnaire. This occurs when people cannot be reached at home either by phone or in-person, when they refuse to be interviewed, 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.[31] With the exception of questions that ask for the respondent's income, item nonresponse rates exceeding 5 to 7 percent are rare in ordinary surveys (Craig and McCann 1978). In CV surveys, however, nonresponse rates of 20 to 30 percent for the WTP elicitation questions are not uncommon when: (1) the sample is random and therefore includes people of all educational 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 percent 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 someone offer an unconsidered guess at an amount or having him say he does not know how much it is worth to him, the latter behavior is preferable, provided appropriate procedures to compensate for the resulting item nonresponse are used.

Both unit and item nonresponse result 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. For example, if 1,000 households are selected by probability-based methods for a CV sample, and valid WTP amounts are obtained for only 800 of these households, the researcher has to determine what effect the missing 200 households have on the WTP estimate. Put another way, can 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 1000 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),[32] 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 a respondent's refusal is often associated with a lack of interest in the topic of the survey (Stephens and

[31] Item nonresponses on WTP questions fall into four general categories: (1) don't knows, (2) refusals, (3) protest zeros, and (4) responses which fail to meet an edit for minimal consistency.

[32] 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 are discussed in detail in Mitchell and Carson (1989: Appendix C).

Hall 1983). Therefore, it seems reasonable to assume that people who are less interested in the amenity will value it differently than will 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 characteristics of these subgroups.[33]

To determine whether observed nonresponse results in bias for a given study, two questions need to be addressed. One question is whether there are differential response rates across identifiable categories or groups of households — for example, users versus nonusers, different educational levels, and so forth — and the other is whether there are systematic differences 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.[34] Sample nonresponse bias will occur if, for example, the sample underrepresents the proportion of low-income households in the population, and these households hold different WTP amounts for the amenity than do households of other income levels. Even if the proportion of low-income households in a study's sample were representative, the study could still suffer from sample selection bias if somehow — either by differential selection or by a higher rate of item nonresponse once interviewed — the low-income people who gave usable WTP amounts differed in their preferences for the good from those low-income people who did not.[35]

The in-person, telephone, and mail survey methods have different vulnerabilities to the sample nonresponse and selection biases. But mail surveys are particularly prone to errors from these sources, especially the latter. This occurs because the unit response rates for mail surveys are lower than those for phone or in-person surveys. Also, the potential for sample selection bias is higher because the questionnaires are self-administered. In this situation, researchers lack control over the process of receiving 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

[33] As are other types of survey variables (Kalton 1983).

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

[35] 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 = $f(X,\beta) + U$ where $f(X\beta)$ 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.

who are too incompetent to be interviewed 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 the specific topic of the survey is made known to them.[36] Third, studies of people who refuse personal or telephone interviews suggest that refusals occur because of general rather than survey-specific reasons (Stinchcombe, Jones, and Sheatsley 1981; T. W. Smith 1983).

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 or not to respond, including the decision to lay it aside, is likely to be influenced by his or her 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 likely the respondent is to fill it out and send it back (Heberlein and Baumgartner 1978; Tull and Hawkins 1984).[37] Because in the case of public goods the respondent's interest in the subject matter is likely to correlate with the value the good has to the respondent, there is a likelihood that nonrespondents will hold lower or even \$0 values for the good compared with respondents of equivalent demographic categories. In short, mail surveys have a strong potential for sample selection bias, which suggests that information from those who happen to give valid WTP answers cannot be used to infer or to impute WTP values for the nonrespondents.[38] This is one of the reasons market research texts (e.g., Tull and Hawkins 1984) do not recommend their use for general populations.[39]

Richard Bishop has suggested putting in zeros for nonresponses to mail surveys as a conservative assumption that also encourages agencies to fund extensive efforts to get high response rates. Almost no completed survey will represent a simple random sample of the population of interest. When using the results of the survey to make estimates, the effects of stratification and cluster, which appear in the best full probability samples, should be taken into account. Weighting to correct for sample nonresponse should also be taken into account. Imputation should be done for item nonresponse and corrections should be made for sample selection bias. This attention to sampling and response issues is extremely important and often strongly influences results.

[36] This presumes, as is the case with many surveys, that the interview topic is described in general terms when the respondents' cooperation is first requested to avoid this type of bias. For example, the interviewer would say they 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."

[37] 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.

[38] For a discussion of the techniques available to compensate for bias due to nonresponse see Mitchell and Carson (1989).

[39] 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 et al. (1980). The Wellman et al. study compared early and late respondents with a mail non-CV outdoor recreation survey that achieved a 70 percent 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 follow-ups to increase recreation survey response rates might better be expended on other phases of the research process." This finding is an insufficient basis to assume random nonresponse as Wellman et al. did not study the 30 percent of their sample who failed to respond to their survey. There are no grounds for believing that late respondents to mail surveys such as theirs are a valid surrogate for the nonrespondents; there is *a priori* and empirical (Anderson, Basilevsky, and Hum 1983) evidence to the contrary.

5.7 Family of α-Trimmed Means

The family of α-trimmed mean estimators drops the α largest and α smallest observations and then calculates the mean value of the remaining observations. The mean is the extreme case where α equals zero and the median is the other extreme where α is 50 percent. For a large class of symmetric distributions, the maximum likelihood estimator can be written in terms of α going from zero to 0.5 as the tails of the distribution become "fatter." Constructed market data tends to be characterized by thick-tailed distributions. These distributions appear to become increasingly asymmetric as the mean willingness to pay becomes larger and more closely tied to the participant's income level — a finding which should not be too surprising. It is a finding, though, which forces the researcher or the policy maker to choose an α. A good way to display the implications is to display a table of the α-trimmed means for different α and to do the benefit-cost analysis using each of these values. Note that because for all individuals i, WTP_i must be nonnegative, the left-hand side outliers are constrained to be zero so that the use of any positive α will typically reduce the estimate of mean WTP.

Even if mean WTP is the desired statistic, using a small nonzero α value may be appropriate, particularly if a hypothetical rather than simulated market is being used. For mean WTP, the main difference between behavior in a hypothetical versus a simulated market appears to be that participants in a hypothetical market take the exercise less seriously than those in a simulated market; however, this does not appear to be the case for WTA markets. For WTA markets, Bishop and Heberlein (1979; 1986) found large differences between hypothetical and simulated markets. This usually manifests itself in mean WTP having a large standard error. Examination of the data usually exhibits a number of implausible large outliers. Use of an α of 0.05 or 0.1 will eliminate the dominant influence of these observations.

5.8 Experimental Design

As researchers have gained more experience with constructed markets, experimental design has taken on a more important role. This is due, in part, to the increasing recognition that many of the early experiments had low power and, in part, to the increasing cost of doing experiments, particularly experiments involving simulated markets in which a considerable amount of money is at stake. The experiments being performed are also taking on new complexity as the hypotheses being tested become more complex.

In designing an experiment involving any type of constructed market, the researcher first needs a clear null hypothesis to be tested and alternatives. Designing an experiment with a clean test between two well-defined specific alternatives is very difficult. Drawing conclusions from rejecting the null hypothesis if more than one possible specific alternative exists, is always dangerous. Bishop and Heberlein's (1979) goose hunting experiment is one of the best known examples. In their study, WTP from a contingent valuation experiment was much smaller than WTA from a simulated market experiment. Bishop and Heberlien concluded, and this was accepted by most contingent valuation researchers, that CV WTP underestimated true WTP since true WTP and true WTA were, according to Willig's results, supposed to be close, and the simulated WTA was accepted as a good estimate of true WTA. Bishop and Heberlein had been unable, for legal reasons at the time, to conduct the simulated WTP experiment. Their later results (1986) showed the simulated WTP and CV WTP were close but quite different from both simulated and CV WTA.

Two other typical problems with constructed market experiments exist. The first is the lack of random assignment of participants to treatments. This usually occurs when two populations are presumed to be similar so that the treatment effect is confounded with the two populations. The second, already alluded to, is the lack of statistical power. By this I mean that in many CV experiments the treatment effect would have to be so large for the null hypothesis to be rejected that for all practical purposes the test is meaningless. Then worse, the failure to reject the null hypothesis may lead the researchers to conclude that the effect is not present. Constructed market experiments are particularly prone to a lack of statistical power due to the large coefficients of variation typical of this type of data and due to the presence of a significant number of outliers.[40] Mitchell and Carson (1989) provided a lengthy appendix on designing experiments that recommends, among other things, a test on medians instead of means (due to the much smaller coefficients of variation and hence smaller sample size needed for a given level of power) and the use of nonparametric statistical tests that are less sensitive to outliers.

[40] The coefficient of variation is the standard deviation divided by the mean. For constructed market data, the coefficient of variation is typically greater than one, which is quite large by experimental standards but is reflective of the degree of income variation in the United States.

5.9 Estimation of Valuation Functions

For many environmental amenities, such as air quality and water quality, the economic question the policy analyst is often asked is: What are the benefits of improving the quality level from A to B when level B is assumed to be preferred to A? There are two ways this question can be answered in a constructed market framework. The first is simply to ask a respondent what he or she is willing to pay to have the quality level rise from A to B. The second is to estimate a valuation function that describes willingness to pay for marginal changes in the quality level. The advantage of the first approach is that the analyst does not have to make assumptions about the form of the utility or the willingness-to-pay function. The first approach's disadvantage, of course, is that it is not very informative on changes other than from A to B, except possibly as an upper or lower bound. The valuation function approach has the opposite advantages and drawbacks. The need to estimate the benefits of a change other than A to B or the desire to trace out a large part of the total or marginal benefits curve leads researchers in the direction of estimating valuation functions.

Estimation of a valuation function raises a number of issues. These issues can be divided into two groups.[41] The first group concerns statistical issues; the second concerns economic issues. The statistical issues revolve around how to optimally estimate the region of the response surface — that is, the benefits curve — in which the researcher is most interested. This problem can be thought of as a special type of experimental design. The economic issues revolve around which, if any, restrictions to impose on the utility or willingness-to-pay function, and which, if any, characteristics of individual respondents to consider. Often, distinctions between the statistical and economic issues become blurred.

Statistically, one wants to estimate the relationship:

$$\text{WTP} = f(\text{environmental quality level}). \tag{5.5}$$

Clearly, the more quality levels that one asks for, the more flexible is the form for $f(\cdot)$ that can be supported by the data. For instance, if only two quality levels are asked about, the researcher can only fit a straight line or a curve with a constant elasticity. Thus, to allow for the possibility of a different curvature, the researcher must either ask individual respondents about more levels of the good or increase the sample size and ask the additional respondents about different levels. The choice of the quality levels will also influence what

[41] For simplicity, it was assumed that the elicitation method has already been chosen and hence whether the data will be of the continuous or discrete type. Of course, the requirements of estimating a valuation function may influence the elicitation method chosen. In particular, the amount of information in discrete responses is substantially less than that in continuous responses, thus making the task of estimating a reliable valuation function with discrete data more difficult.

can be estimated. If two quality levels very close together are chosen, then it is likely that the WTP function will appear linear. One of the best guides to choosing optimal levels is to inquire about levels just above and below the range defined by existing levels and likely policy options. Box and Draper (1987) provide a good guide to response surface estimation.

The specification in (5.5) can be enriched by the incorporation of covariates. There are two reasons for doing so. The first is to increase the statistical efficiency by reducing the unexplained variance. The second is to test whether or not WTP appears to be driven by predictable factors, particularly those suggested by economic theory. If researchers randomly assign subsamples to different quality levels or if they ask each individual about each quality level and then stack the observations, the quality level will be orthogonal to the individual's characteristics.[42] Estimation of the model with covariates, of course, raises the issue of consistency with utility theory and the issue of whether utility theory imposes any restrictions on the model which should be tested.

A couple of other issues should be raised when considering the estimation of a valuation function. The first is how to treat protest zero responses. The approach used most often is to discard them. This is clearly wrong from a statistical point of view. A better approach is to explicitly model them using some type of maximum likelihood or nonparametric framework. Another problem with constructed market data is the presence of outliers (usually on the right side). Again the typical course of action has been to discard them. Robust regression techniques that down weight these outliers seems to be a better approach and one much more justifiable on statistical grounds.

5.10 Open Issues

While many of the fundamental issues in constructed markets are now settled, there are, nonetheless, a number of open issues with respect to constructed markets. These fall into four main categories: (1) the use of constructed markets in new application areas; (2) the role of information in constructed markets; (3) the exploration of theoretical issues using constructed markets; and (4) the statistical issues in the design and analysis of constructed markets.

One logical way of depicting the history of constructed markets is in terms of the process by which researchers determined how to use constructed markets to value a particular environmental amenity. Perhaps the best example is the long chain of air quality studies that started with Randall, Ives, and Eastman (1974). The main focuses of these studies was how to portray changes in air

[42] If the individual is asked about several levels (and those observations stacked for the purposes of estimation) then it may be reasonably expected that there is a panel data type correlation structure induced.

quality to participants and how to define a market structure for air quality. Each new study produced insights into what participants thought they were buying. Occasionally, there was a major advance or failure in describing air quality or the market in which it was sold. Now, a researcher desiring to do an air quality study in a different location has a firm foundation upon which to start. Each new environmental amenity produces a new challenge to researchers. They must determine how to describe it to participants, why the participants want it, and what reservations the participants may have about a program to supply it. This is a new experience to economists who generally have been able to ignore what actually motivates someone to purchase a good.

One of the most exciting new areas for the use of constructed markets is valuing risk reductions from environmental pollutants (e.g., Smith, Desvousges, and Freeman 1985). Psychologists have long argued that changes in low-level risk are very difficult for people to understand. Researchers have been experimenting with a number of different ways of expressing risks and are enjoying some success. Work is currently being conducted on risk from groundwater contaminants, pesticides, and radon. Another new area receiving considerable attention is natural resource damage assessment.[43] Natural resource damage assessment creates a host of new problems because the damage usually has already occurred so that it is difficult to obtain an *ex ante* welfare measurement, and because there is usually an easily identifiable "guilty" party thus creating the clear opportunity for strategic behavior that is usually lacking in most contingent valuation studies.

If a researcher accepts the argument that the values obtained in a constructed market exercise are contingent on the information available to participants, then a systematic exploration of how information influences values would appear to be necessary. What would be ideal is a quantification of how various types of information influence WTP responses, in particular, an investigation into the role of uncertainty with respect to likelihood of the amenity actually being supplied and into the role of the agent receiving payment for the amenity.

Constructed markets allow researchers to test a number of fundamental issues related to economic theory. This has been long recognized by experimental economists using simulated markets. With the exception of testing the relationship between WTP and WTA, contingent valuation has been less used for this purpose.[44] Other areas in which constructed markets should be useful are in examining how people actually discount future environmental amenities,

[43] See, for instance, Carson and Navarro (1988), Mitchell and Carson (1988), and Schulze (1988).

[44] In part this is due to the strong suspicion that economists have with regard to responses to hypothetical survey questions. The large differences between WTP and WTA consistently found in contingent valuation studies was ascribed to the hypothetical nature of the questions until Bishop and Heberlein's (1979, 1986) simulated market studies began to show the same large differences.

such as risk reductions (Horowitz and Carson 1988), and how to transfer the values obtained in one constructed market study to a new situation where a benefit estimate is needed.[45] The issue raised by Hoehn and Randall (1989) of aggregating benefits across geographic areas and across policies is still largely unexplored.

While the success of contingent valuation has largely exceeded the expectations of its early proponents, one of their great hopes for contingent valuation was that it would provide a cheap alternative to the other benefit measurement techniques. Unfortunately, contingent valuation has not proven cheap to implement. In order to minimize cost for a specified level of precision, contingent valuation researchers are starting to examine whether it is possible to use more efficient sampling plans and experimental designs. Contingent valuation data, in large part because it is survey data, is also not as clean as the macro or financial data with which economists typically work. This feature of the data is leading contingent valuation researchers to look at techniques for handling outliers and missing data and the implications of using those techniques. The shift to discrete choice contingent valuation questions has focused attention on discrete choice estimators. The ability to frame questions in particular ways is giving insight into what the discrete choice question is measuring (Cameron and James 1987) and can be exploited to gain more efficient estimates of willingness to pay.

[45] To date there has been little work done on this topic. Smith and Kaoru (1988) have undertaken the first formal study of benefit transfer but have focused on recreational demand travel cost studies rather than contingent valuation studies. Carson and Mitchell (1988) showed how Smith and Desvousges's (1986b) Mongahella River water quality CV estimate could be obtained from their CV study of national water quality benefits.