Working Paper Series

WORKING PAPER NO. 697

CONTINGENT VALUATION AND ECONOMICS

by

W. Michael Hanemann

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS
BERKELEY
CALIFORNIA AGRICULTURAL EXPERIMENT STATION

University of California
DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS AND POLICY
DIVISION OF AGRICULTURE AND NATURAL RESOURCES
UNIVERSITY OF CALIFORNIA AT BERKELEY

WORKING PAPER NO. 697

CONTINGENT VALUATION AND ECONOMICS

by

W. Michael Hanemann

California Agricultural Experiment Station
Giannini Foundation of Agricultural Economics
March, 1994
1. Introduction

The question at issue is this: can we, as social scientists, measure in monetary terms the value that people place on the natural environment, or other public goods, when there is no other corroborating market behavior that would allow measurement by revealed preference methods? Are we forced to remain silent when methods such as travel cost or hedonic pricing [Freeman (1993)] are unavailable? Obviously, it would be simpler if we could restrict economic analysis to commodities where the conventional techniques of revealed preference can be applied. But, this would be a truly procrustean solution: it would put beyond the bounds of analysis many commodities that people value but do not purchase through a market or that they value in part for reasons unconnected with their own purchase and use.

One of the great errors of analysis is the fallacy of misplaced precision—measuring the wrong thing with exquisite precision, rather than the right thing with lesser accuracy. The need to avoid this error has been recognized since the very beginnings of cost-benefit analysis in the United States. In 1934, when a National Resources Board committee was appointed to consider cost-sharing for water resource projects, its report pointed out the need to study "the part played by intangible factors" in assessing public works programs. There was similar consideration in the Flood Control Act of 1936, which permitted federal funding of flood control projects "if the benefits to whomsoever they may accrue are in excess of the estimated costs, and if the lives and social security of people are otherwise adversely affected." As this formula was applied over the next decade to other public investment programs, there was a general recognition of the legitimate role of intangible factors such as national defense, saving human lives, and recreational or aesthetic impacts. This concern was endorsed in the first bible of cost-benefit analysis, the U.S. Inter-Agency River Basin Committee's so-called Green Book, published in 1950. This explicitly recognized the category of intangibles and prescribed that these "need to be described with care and should not be overlooked or minimized merely because they do not yield to dollar evaluation."

When the Green Book was being written, economic valuation was generally perceived in terms of market prices. To value something, one had to ascertain an appropriate market price, adjust it for market imperfections, and then multiply this price by a quantity. There was no possibility of valuing items, such

---

1 I alone am responsible for the views expressed here. Most of what I know about CV has been learned from working with Richard Carson, Robert Mitchell, and Kerry Smith. I have also gained much from conversations and collaboration with Richard Bishop, Gardner Brown, Michael Conaway, Barbara Kanninen, Ray Kopp, Bengt Kristrom, Jon Krosnick, John Loomis, Stanley Presser, Alan Randall, and Paul Ruud. I am greatly indebted to them. I thank Nicholas Flores, Sandra Hoffmann and Jessica Wooley for their excellent assistance in preparing this paper.
as intangibles, for which no market price existed. Two developments changed this view. The first was the recognition, prompted by the "new welfare economics" of the 1940s and especially Hotelling's paper on public utility pricing, that the appropriate welfare criterion is maximization of aggregate consumers' plus producers' surplus. Hence, while market prices can safely be used to value marginal changes in the supply or demand for market commodities, the impact of non-marginal changes is measured by the change in areas under demand and supply curves. The second development was Samuelson's analysis of public goods and his finding that their valuation must be based on vertical aggregation of individual demand curves. Together, these developments led to an important paradigm shift during the 1950s—one that contributed directly to the emergence of non-market valuation and is still being felt in the current debate on contingent valuation (CV). This shift changed the focus of valuation away from market prices towards demand and supply functions as the underlying repositories of value. These functions are behavioral relations, and the implication of the paradigm shift was that economics is not just the study of markets but more generally the study of human preferences and behavior.  

The conceptual link to non-market valuation is the recognition that, while a demand curve is not observable if there is no market for a commodity, there still exists a latent demand curve that perhaps can be teased out through other means. The classic example is a public bridge. Despite the absence of a charge, there still is a demand curve for the bridge and the bridge's value is given by the area under this curve. The measurement problem is to uncover the latent demand curve. The two main approaches to non-market valuation—the travel cost method and CV—represent alternative attempts at a solution. 

Remarkably, both approaches were first proposed in 1947. The travel cost method was suggested by Harold Hotelling in a letter to the National Park Service (NPS) responding to their request for advice on how to measure the economic benefits from the national parks when there were no revenues from entrance fees. Hotelling saw the similarity with valuing a bridge. Indeed, he felt that valuing a park was easier. Other inputs were involved in using the park, such as expenditures for travel, lodging and equipment: these expenditures were not captured by the NPS but, still, they set a price on visiting the park. Moreover, this price would vary among visitors coming from different origins. By measuring this price for visitors to a site one could estimate their demand curve, and then calculate consumer's surplus in the usual manner. This was not followed up until Trice and Wood (1958) and Clawson (1959). Within five years, 

---

2 One might also say that, under the new view, markets are of interest to economists because they generate choices rather than just prices. In this regard, Diamond and Hausman's (1994) insistence that data on market transactions are the touchstone for appraisal of economic analysis seems strange. Economics owes its prestige among the social sciences to the fact that it has broadened its focus beyond the marketplace to other forms of human behavior such as political markets (Downs) and the family (Becker).

3 The third approach—hedonic pricing—aims more at marginal valuation, since the underlying market demand function generally cannot be recovered from the hedonic price equation [Epple (1987)].
the travel cost method was well established in the literature.

Also in 1947, S. V. Ciriacy-Wantrup published a paper on the economics of soil conservation in which he observed that some of the benefits were not conventional market commodities. He suggested the following solution to what he saw as the central problem, obtaining a demand schedule for such goods: "[Individuals] may be asked how much money they are willing to pay for successive additional quantities of a collective extra-market good. The choices offered relate to quantities consumed by all members of a social group. ... If every individual of the whole social group is interrogated, all individual values (not quantities) are aggregated. The results correspond to a market-demand schedule." While noting the possible objection that "expectations of the incidence of costs in the form of taxes will bias the responses to interrogation," he felt that "through proper education and proper design of questionnaires or interviews it would seem possible to keep this potential bias small." As noted in Portney (1994), the first application was by Davis (1963) in his Harvard Ph.D dissertation on the economic value of recreation in the Maine woods. Within ten years, CV studies were becoming commonplace, and CV was established as a technique for nonmarket valuation [e.g., Maler (1974)].

Clearly, travel cost was accepted more quickly than CV. One reason is that it was much closer to the standard paradigm of market commodities. Most economists not only believed in revealed preference but considered the purchase of market goods to be the only valid expression of preference; voting, a paradigm Maler and Schelling (1968) linked with CV, was of interest to few economists. A second was that CV requires survey research expertise; it cannot be done well by economists alone. It took until the 1980s for adequate links to be forged with the other social sciences. Like CV, the travel cost method was not accepted without debate. Eckstein (1958), for example, strongly opposed it as inaccurate and likely to harm the credibility of benefit-cost analysis in general. Furthermore, there was a distinct learning curve in applying travel cost—the first studies in the 1960s were crude and unsophisticated, but the methodology was steadily refined over the next 15 years. The same was true of CV during the 1980s.

This history gives perspective to the current debate. The positions being staked out today on CV are not new—the contours of the debate have been known for some time. In 1958, Eckstein made the case

---

4 This anticipates Samuelson's result about vertical instead of horizontal aggregation of individual demands for a collective good. The early history of CV and travel cost is described more fully in Hanemann (1992).

5 The need for these links was first stressed in Cummings et al. (1986) and is a major theme of Mitchell and Carson (1989), now the standard work on CV, which places it in the broader context of social science research, combining elements of sociology, psychology, political science and market research as well as economics.

6 From the beginning there was an awareness of certain problems with travel cost still being studied today, such as measuring price, dealing with time, multiple destinations, substitute sites, and site quality. Similarly with CV. Indeed, Knetsch and Davis (1966), wrote a paper setting out the problems with the two methods. They made an empirical comparison of WTP estimates for forest recreation in Maine using both methods, and found a close match.
against CV (which he liked no better than travel cost).  

To determine the price that consumers would be willing to pay if they could purchase national defense, they would have to be interviewed. Since they know perfectly well that the defense budget will not be influenced by their response, and in view of the unpleasant tax possibilities of giving a higher answer, they would have no reason to give truthful replies. And even if they were perfectly guileless ... their lack of experience in purchasing this 'commodity' would result in answers which might be far different from the answers they would give if actual purchases were involved. 

In 1968, Schelling framed the case for CV:

There are two main ways of finding out whether some economic benefits are worth the costs. One is to use the price system as a test of what something is worth to people who have to pay for it. ... Another way ... is to ask people. This can be done by election, interview or questionnaire ... Like the price system, these methods may be ambiguous. It is sometimes argued that asking people is a poor way to find out, because they have no incentive to tell the truth. That is an important point, but hardly decisive. It is also argued, and validly, that people are poor at answering hypothetical questions ... While this argument casts suspicion on what one finds out by asking questions, it casts suspicion too on those market decisions that involve remote and improbable events. ... This problem of coping, as a consumer, with increments in the risk of unexpected death is very much the problem of coping with hypothetical questions, whether in response to survey research or to the man who sells lightning-rod attachments for the TV antenna... In any case, relying exclusively on market valuations and denying the value of direct enquirey in the determination of government programs ... would depend on there being, for every potential government service, a close substitute available in the market at a comparable price. It would be hard to deduce from first principles that this is bound to be the case. 

What, then, has changed since these words were written? Against CV, very little has changed. Professors Diamond and Hausman (DH) and their colleagues have conducted some experiments that appear to support some of Eckstein's concerns. If examined carefully, however, their evidence is less than convincing. For CV, there have been two changes. First, there is now less concern about free-riding—the empirical evidence shows that it is considerably less ubiquitous than feared in the 1960s [Davis and Holt (1993)]. Second, there was a steady improvement in CV methods during the 1980s, making it more reliable and robust. The CV literature now contains more than 1500 studies and papers from over 40 countries covering a wide range of topics—transportation, sanitation, health, the arts, education, the environment [Carson, et al. (1993), Navrud (1992)].

So, why all the sturm und drang about CV at this time? The answer, as Portney (1994) explains, is

---

7 National defense, like any other broad aggregate, is a poor commodity to value through CV. Just what does it mean to have zero national defense, and would respondents find this a credible alternative? But, I have no doubt that one could use CV to value specific military programs—just as the Swiss sometimes hold referenda on them ("Swiss Voters Approve $2.5B Order of McDonnell-Douglas F/A-18 Fighters." Wall Street Journal 6/7/93).

8 Cambridge Economics Associates Inc. (1992) is a set of papers funded by the Exxon Company and presented at a conference in Washington, DC on April 2-3, 1992. These appear with minor changes in Hausman (1993). Since I wrote most of this paper without having seen Diamond and Hausman (1994), I relied heavily on those two sources for a statement of their position, especially their chapter in Hausman (1993), to which I henceforth refer as DH.
its use in litigation over damages from oil spills, especially litigation arising from the *Exxon Valdez* oil spill in March, 1989. In this paper I focus on the general question of how one measures preferences for the environment. I will not discuss the *Valdez* litigation or other aspects of natural resource damages.

The rest of this paper is organized as follows. Section 2 deals with the economic theory of nonuse value and shows why it cannot be measured by revealed preference methods. Section 3 describes aspects of CV practice that enhance its reliability. Sections 4 and 5 compare CV with public opinion polls and market research surveys. Section 6 reviews the empirical evidence marshalled by Hausman (1993), and Section 7 addresses economic theory and CV, including DH’s theoretical arguments about CV.

2. The Economics of Nonuse Value

People have long valued wildlife for many reasons besides wanting to shoot, eat, wear, or otherwise utilize it. The number who do so has grown steadily [Hoage (1989)]. The first systematic attempts to account for this in economic terms were by Weisbrod (1964) and Krutilla (1967). Weisbrod focused on uncertainty and what became known as "option value"—e.g., some people who do not now visit a wilderness area have a positive WTP to protect it in order to preserve their option of visiting it in the future. Krutilla focused on "bequest value"—some people have a positive WTP for the wilderness area because they want to preserve it for future generations—and on what became known as "existence," "nonuse" or, more recently, "passive use" value—some people have a positive WTP for the area, even if neither they nor their children would ever visit it. As Krutilla put it, they may "obtain satisfaction from mere knowledge that part of the wilderness in North America remains, even though they would be appalled by the prospect of being exposed to it." These demands to protect the resource are not reflected in market prices but, Weisbrod and Krutilla contended, they do belong in a social benefit-cost analysis.

These arguments are now widely accepted. However, because they were developed in a verbal manner, there remained some ambiguity over the precise definition of the concepts that was not resolved until the 1980s. I focus here on nonuse value since this underlies the case for CV. A starting point is the welfare theory associated with the travel cost model. This was first articulated by Maler (1974), who sought to extend standard welfare theory of price changes to changes in the supply of a public good. His model is a version of Lancaster's (1966) model of demand with characteristics. In this model, the consumer has preferences over conventional market commodities, denoted by the vector $x$, and over a set of other items denoted by the vector $q$. Whether the elements of $q$ are attributes of private market goods, supplies of public goods, or amenities, the key is that the consumer views them as exogenous. He can freely vary the

---

9 I was one of a group of economists retained by the State of Alaska in connection with its suit for natural resource damages; our CV analysis is presented in Carson et al. (1992).
x's, but not the q's. Preferences are represented by a utility function, \( u(x,q) \), which is nondecreasing in \( x \) and \( q \), and strictly quasi-concave in \( x \) but not necessarily \( q \). The individual maximizes \( u(x,q) \) with respect to \( x \) subject to the usual budget constraint, which yields ordinary demand functions for the market goods, \( x_i = h_i(p,q,y) \), \( i = 1,\ldots,N \), and an indirect utility function, \( v(p,q,y) = u[h(p,q,y),q] \), where \( p \) represents prices of the x's and \( y \) is the consumer's income.\(^{10}\) Dual to this are the compensated demand functions \( x_i = g_i(p,q,u) \), \( i = 1,\ldots,N \), and the expenditure function \( m(p,q,u) \). Suppose, that \( q \) increases from \( q^0 \) to \( q^1 > q^0 \) while prices and income remain constant at \( (p,y) \). Accordingly, the individual's utility increases from \( u^0 = v(p,q^0,y) \) to \( u^1 = v(p,q^1,y) \). By analogy with the welfare theory of price changes, Maler defined compensating and equivalent variation measures for this utility change. The compensating variation, \( C \), which corresponds to the individual's WTP for the change, satisfies

\[
(1) \quad v(p, q^1, y - C) = v(p,q^0,y).
\]

The key question is how this quantity is measured. Maler developed a method based on what he called weak complementarity. This arises when there is some private market good, or group of goods, such that, when those goods are not being consumed, the marginal utility \( \partial u/\partial q_k \) is zero for all the elements of \( q \) that change between \( q^0 \) and \( q^1 \). For example, \( q \) is the availability of freeways, ownership of an automobile is a complementary good, and, without it you don't care whether or not there are freeways. Or, \( q \) is water quality at some beach, the complementary good is visits to that beach, and you don't care whether the water is clean if you don't visit the beach. Whether this property holds is an empirical question. Suppose that \( x_i \) is inessential and weakly complementary with the set of \( q \)'s which change, so that \( x_i = 0 \Rightarrow \partial u/\partial q_k = 0 \). Maler showed that the quantity \( C \) is related to the change in the area under the compensated demand curves for the weakly complementary commodity when evaluated at \( q^0 \) and \( q^1 \):

\[
(2) \quad C = \int_{p_1^*}^{p_1^*} [g^1(p, q^1, u^0) - g^1(p,q^0, u^0)] \, dp_1,
\]

where \( p_1^* \) is the cut-off price such that \( g^1(p_1^*, p_2,\ldots,p_N, q^0, u^0) = 0 \), \( t = 0,1 \).

In practice, the usefulness of this result is limited by the need to know the compensated demand functions, as opposed to the ordinary demand functions. As an approximation, one might consider using areas under ordinary demand functions —i.e., substituting \( h^i(p, q^i, y) \) for \( g^i(p,q^i,u) \) in (2)—but the resulting measure does not necessarily have the same sign as \( C \), let alone the same magnitude. Instead, one can

\(^{10}\) This model includes the household production model as a special case when the \( q \)'s are production function parameters. It is discussed further and contrasted with the hedonic demand model in Hanemann (1982).
apply Hurwicz and Uzawa’s (1971) integrability results and solve the fundamental system of partial
differential equations

\[
\frac{\partial m(p,q,u)}{\partial p_i} = h'[p,q,m(p,q,u)] \quad i = 1, \ldots, N,
\]

for the expenditure function \(m(p,q,u)\), from which compensated demand functions can be derived.\(^{11}\) With
this approach, moreover, one can by-pass (2) and calculate \(C\) directly from the indirect utility function on
the basis of (1). Weak complementarity per se is not needed.

With regard to nonuse value, the definition due to McConnell (1983) and Hanemann (1988) rests on
the particular structure of the utility function. Suppose that the direct utility function takes the form

\[
u(x,q) = T[\bar{u}(x,q), q]
\]

for some functions \(T[\cdot]\) and \(\bar{u}(\cdot)\), where \(T[\cdot]\) is increasing in both arguments and \(\bar{u}(\cdot)\) is nondecreasing in
both sets of arguments and strictly quasi-concave in \(x\). The key to (4) is that the marginal rate of
substitution among the \(x\)'s is independent of \(T[\cdot]\); it depends on \(q\) to the extent that \(q\) enters into \(\bar{u}(\cdot)\), but
not to the extent that \(q\) enters into \(T[\cdot]\). Let \(\bar{v}(p,q,y)\) be the indirect utility function associated with \(\bar{u}(x,q)\).
Given (4), the compensating variation for the change in \(q\), \(C\) from (1), can now be expressed as a sum

\[
C = \bar{C} + C^T,
\]

where \(\bar{C}\) satisfies \(\bar{v}(p,q^1,q^1-y-C) = \bar{v}(p,q^0,y)\), and \(C^T\) satisfies \(T[\bar{v}(p,q^1-y-C^T),q^1] = T[\bar{v}(p,q^1,y),q^0]\). \(\bar{C}\) is the
individual's WTP for the change in \(q\) based on \(\bar{u}(x,q)\), while \(C^T\) is the balance of \(C\) once \(T[\cdot]\) is taken into
account. \(\bar{C}\) is the natural candidate for use value—it reflects that part of preference for \(q\) associated with
consumption of private market goods. \(C^T\) is the natural candidate for nonuse value—it stems from that
preference for \(q\) which is separate from preference for the \(x\)'s. Nonuse values arise, then, whenever the
utility function can be factored into the form of (4). Hanemann (1988) showed that the absence of weak
complementarity is neither necessary nor sufficient for this to occur.

Two consequences follow from this definition. Both \(u(x,q)\) and \(\bar{u}(x,q)\) lead to exactly the same
ordinary demand functions, \(h'(p,q,y)\), even though they generate different compensated demand functions.

\(^{11}\) Maier and Willig (1976 a,b) pointed out that this system of differential equations provides a practical way of
recovering the utility function; see also Hanemann (1980) and Hausman (1981). LaFrance and Hanemann (1989)
provide a fuller account of what can and cannot be recovered from (3). In this paper, I use the term "travel cost"
generically for methods that recover \(v(p,q,y)\) from (3) for any commodity, not just recreation.
It follows that the Hurwicz-Uzawa integrability approach based on (3) will recover \( \bar{v}(p,q,y) \) but not \( T[.] \). As a result, revealed preference methods cannot be used to recover \( C_T \). In contrast, contingent valuation does recover \( C = \bar{C} + C_T \). Secondly, any \( q \) whose quantity the individual takes as fixed may generate nonuse values; this depends on the structure of preferences, not on the type of change in \( q \). It is an empirical question, therefore, whether nonuse values arise or how large they are. Attempts to determine this \textit{a priori} have no foundation in economic theory.

DH invoke economic theory repeatedly, claiming that it is not compatible with the empirical CV results in Hausman (1993). But, they are incorrect on the theory. Suppose that only one of the \( q \)'s changes [e.g., \( q^0 = (q_1^0,q_2^0,...,q_k^0) \) and \( q^1 = (q_1^1,q_2^1,...,q_k^1) \)]. From (1), the value of \( C \) depends on the remaining \( q \)'s; from (4), so do the values of \( \bar{C} \) and \( C_T \). Madden (1991) shows that an increase in \( q_2 \), say, will \textit{lower} \( C \) if \( q_1 \) and \( q_2 \) are what he calls R-substitutes, and \textit{raise} it if they are R-complements. Likewise, \( C \) depends on \( y \); while the income elasticity of \( C \) is related to the income elasticity of demand for \( q_1 \) and has the same \textit{sign}, the two elasticities are not necessarily similar in magnitude, as DH (p. 42) assume. The income elasticity of \( C \) depends on not only the income elasticity of demand for \( q_1 \) but also the elasticity of substitution between \( q_1 \) and the \( x \)'s. The latter is the more influential in determining the magnitude of \( C \) and can explain large divergences between WTP and WTA [Hanemann (1991a)].

While DH consistently downplay substitution effects between \( q_1 \) and other \( q \)'s or \( x \)'s, these have important economic implications. For example, R-substitution explains the sub-additivity of WTP—if \( q_1 \) and \( q_2 \) are R-substitutes, the WTP for a change in both \( q_1 \) and \( q_2 \) is less than the sum of the WTPs for the changes taken separately. Far from being inconsistent with economic preferences (Diamond et al. pp. 48-49), sub-additivity is likely to be the norm; while all goods cannot be R-complements, Madden shows they \textit{can} all be R-substitutes. Similarly, R-substitution explains what Diamond et al. call sequence aggregation and Kahneman and Knetsch (1992) the embedding problem. If \( q \)'s are substitutes, the WTP for a change in \( q_1 \) is \textit{lower} when it comes at the \textit{end} of a sequence of changes in \( q \)'s than at the beginning, while the WTA for the change in \( q_1 \) is \textit{higher} when it comes \textit{later} in a sequence [Carson, Flores and Hanemann (1992)]. In short, it should come as no surprise that the value of one commodity changes when the quantity of another varies, i.e., that WTP depends on the economic context.

3. Survey Design and the Reliability of CV Responses

Likewise, it should come as no surprise that responses to CV surveys depend on the survey context. Surveys are an exercise in communication, and this is always sensitive to nuance and context. Moreover,

\[12\] In natural resource damages where WTA is the relevant welfare measure, this implies that the usual practice of taking the injured resource as the first item in any possible valuation sequence is a conservative procedure.
preferences are inherently context-laden. A basic lesson of the cognitive revolution in psychology is that context determines meaning. Potter (1990) notes "a central principle of encoding" that the same object is encoded differently in different contexts. It would be simpler if consumers disregarded context and had preferences for commodities in the abstract, as Platonic ideals. But, they don't.

I discuss the implications for economic theory in section 7; here I focus on the implications for CV design. Almost all data used in the social sciences are self-reported data obtained through surveys, such as household expenditure surveys, employment surveys, health surveys, and opinion polls. Like everything involving people, surveys are fallible. There is no magic formula for perfect measurement. Procedures can be identified, however, that are likely to generate more reliable results. In the case of CV, certain aspects of study design enhance validity: avoiding self-administered surveys; using a probability sample; building in quality control through the process of instrument development and testing; making the commodity realistic; making the payment credible; providing appropriate information; legitimizing a "no" response; and using debriefing questions to probe respondents' perceptions and motives.

The survey must be formulated around a commodity that captures what one seeks to value. It must be consistent with economic theory. And, it also must be intelligible and meaningful to respondents in the field. A National Academy of Sciences Committee (1986) described how good researchers study human decision making: "They question respondents about specific situations, rather than asking for generalizations. They are sensitive to the dependence of answers on the exact form of the questions. They are aware that behavior in an experimental situation may be different from behavior in real life, and they attempt to provide experimental settings and motivations that are as realistic as possible." Exactly the same applies with CV. Respondents must be confronted with a commodity that is concrete and realistic, so they take it seriously. The vaguer and less specific the scenario, the more likely they are to treat the valuation as merely a symbolic exercise. Likewise, the valuation question should not be overly counterfactual. For example, asking "If it had been possible to prevent the accident, how much would you have paid?" requires people to deny an event that did occur. Rather than something in the past, one should try to value an action that can plausibly be taken now—e.g., a government program to prevent future accidents. This program can be described with sufficient detail and realism for people to believe in it.

This strategy does have drawbacks. Some respondents may regard preventing a future accident as less serious since the real damage was done in the earlier accident. Others may not believe there will be a future accident because people will now be careful. Yet others may believe that, while there is a chance of a future accident, the program won't prevent it. They may value the environment but refuse to pay...

---

13 In all the social sciences, economics included, procedural invariance in measurement is an ideal, not the reality. It is even violated in the physical sciences for several reasons including the problem that no individual scientific hypothesis is conclusively falsifiable (the Duhem-Quine problem) (Cook and Campbell (1979)).
because they view the proposed program as ineffectual. Hence, there can be a gap between what one wants to value and what respondents do value. But, this may not be a serious problem. First, one can try to ensure that the difference lies in a conservative direction. Second, one uses debriefing questions to quantify these effects. The debriefing comes at the end of the survey and asks respondents whether they had completely believed key parts of CV scenario (e.g., Was the damage as bad as described? Did you think you would really have to pay higher taxes if the program went through?), and why they answered the WTP question as they had (e.g., What was it about the program that made you decide to vote for it?). One can thus monitor for the presence of divergent or inappropriate beliefs, measure statistically how these affected respondents’ WTP, and adjust accordingly.

Similarly, the payment description must meet three key requirements. First, the payment should be perceived as linked directly to provision of the commodity. Second, to make it plausible one should specify a particular context. Third, it is best if the payment is viewed as mandatory once a commitment has been made; e.g., if the program is approved one can’t avoid paying because firms will then raise prices or the government will raise taxes. One wants respondents to take the notion of payment very seriously.  

There has been much debate on the paradigm for CV: is it like markets, voting, or charity? DH say the right paradigm is charity. But, most appeals for charity are not designed to elicit one’s maximum WTP—they just ask for some donation. They usually focus on the symbolic aspects of the donation. Unlike voting, they generally don’t identify a specific outcome that follows from one’s action and would not occur otherwise. For these reasons, charity is generally a flawed paradigm for CV. Environmental protection is usually a collective good, and its supply is perceived as arising through collective action. Therefore, as Schelling (1968) noted, voting is the natural paradigm for CV. This was reaffirmed by the NOAA Panel: "The simplest way to approach the problem is to consider a CV survey as essentially a self-contained referendum in which respondents vote on whether to tax themselves for a particular purpose" [Arrow et al. (1993)]. It set aside suggestions that markets are the right paradigm and that CV be limited to commodities which respondents have prior experience buying. As a California voter last November I had no previous experience of voting on school vouchers, but it was not difficult to form

14 There may be "protest zeroes" because respondents think it unfair that they should have to pay. This, too, can be detected through debriefing questions.

15 The two studies DH cite as showing a lack of commitment in CV, Seip and Strand (1992) and Duffield and Patterson (1992), used payment to an environmental charity rather than voting on a government program. Most of the subjects in Seip and Strand’s study who were followed up afterwards said that they had been expressing a willingness to pay for environmental problems generally rather than to support the particular environmental group.

16 DH refer many times to Andreoni’s model which assumes a “warm glow” from the act of giving to charity. However, I know of no empirical evidence that people get a warm glow from voting to raise their own taxes.
an opinion and cast my vote. Prior experience, *per se*, was far less important than adequate information.\(^7\)

Another issue is whether the payment question should be asked as an open- or closed-ended question. Prior to 1985, most CV surveys used an open-ended WTP question, such as "What is the most that you would be willing to pay for...?" Since then, most major CV surveys have used closed-ended questions like "If it costs $x, would you be willing to pay this amount (vote for it)?" I consider the latter far more reliable.\(^8\) People do not know their WTP for most goods, private or public, in the way they know their shirt size. Nor can they discover it by inspecting their utility function, as in the textbooks.\(^9\) Rather, it is revealed to them as the consequence of acts of judgment when they face choices and make decisions. Whether in the market or in voting, these choices usually are discrete: here is an item, it costs $x, will you take it? Therefore, the closed-ended format is more realistic. Further, there is abundant evidence that respondents find open-ended questions more difficult to answer than closed-ended ones.\(^20\) Even if people have experience buying an item and can state an amount which they would be willing to pay, they may find it hard to state the maximum amount. As a result, open-ended responses can understate maximum WTP.\(^21\) Since the maximum WTP is an extremum, errors of cognition tend to fall on the low side. This bias may be reinforced by strategic behavior associated with open-ended questions which leads

\(^7\) A CV questionnaire should present adequate information, in an impartial and balanced manner, before the valuation question is asked. In the "top-down" procedure of Kemp and Maxwell (1993) and Kahneman and Knetsch (1992), the information comes after. Subjects are asked to value some broad item, e.g., "preparedness for disasters." Then, they are told what things make up this item. They are asked their WTP for one of those components. Then, they are told what it comprises, etc. They also are not told at any stage just what is the change in the quantity of an item if they pay. All of this highlights the symbolic aspects of the valuation exercise at the expense of substance.

\(^8\) Bishop and Heberlein (1979) first used this format for CV; the link with utility theory was developed in Hanemann (1984); see also Cameron (1988) and McConnell (1990).

\(^9\) DH seem to take literally the notion that an individual knows his utility function, e.g., in dismissing the responses by participants in Payne and Schkade's (1993) verbal protocol study. I find those responses thoughtful and reasonable; Professor Solow recently said "they sound an awful lot like Bob Solow in the grocery store." Kahneman et al (1993) characterize responses to CV surveys as "constructed values." But, modern neuroscience suggests that all thought is a construct [see Rose (1992), who observes: "brains do not work with *information* in the computer sense, but with *meaning*...Our memories are recreated each time we remember"].

\(^20\) Kristrom (1990), Desvousges et al. (1992), and Soderquist (1992) give evidence of lower participation rates and higher item non-response rates with open-ended questions. The difference in formats has a larger significance. It corresponds to the distinction in psychology between matching (open-ended) and choice (closed-ended) [Tversky, Sattath and Slovic (1988)], and it underlies most of the evidence on preference reversal [Tversky, Slovic and Kahneman (1990)]. These experiments switch between matching and choice, and the interpretation of the outcome as a reversal of preferences rests on the assumption that the two formats ought to produce identical results. This overlooks the cognitive differences and it treats the two formats as equally reliable. By letting the variance of WTP vary with format but keeping the mean the same, Alberini (1992) generates preference reversals like those observed.

\(^21\) These arguments run in the opposite direction with WTA, and explain why open-ended WTA responses may *overstate* individuals' minimum WTA. Taken together, these arguments suggest that open-ended questions can exaggerate differences between WTP and WTA as compared to closed-ended questions.
respondents to state less than their full value. Unlike free-riding, this is clearly supported by experimental evidence. With the closed-ended referendum format, in contrast, "there is no strategic reason for the respondent to do other than answer truthfully" [Arrow et al. (1993)]. Experience also shows that, with open-ended questions, some respondents think about what the item could cost per household rather than what it is worth to them, which may be more. For all these reasons, the closed-ended format is to be preferred. The NOAA Panel agreed; it recommended using closed-ended questions set in a voting context.

The NOAA Panel also saw probability sampling as essential for CV surveys. With regard to survey mode, it considered mail surveys unlikely to be reliable because of low response rates, difficulties with sample frames, and literacy problems among the general population. It also saw lack of control over the interview process as a general problem with self-administered surveys—a respondent can look at all the questions before deciding which, if any, to answer. In-person interviews are expensive but offer practical advantages in permitting graphical aids, maintaining respondent motivation, and monitoring respondent performance. Telephone surveys may be satisfactory if respondents are already adequately informed, or the issue can be conveyed briefly and without visual aids. The Panel concluded that in-person CV surveys would usually be preferable.

4. CV and Voting

The analogy with voting is crucial to the debate on CV. The NOAA Panel saw the analogy as natural. DH disagree vehemently. They attempt to buttress their case by citing Magleby's (1984) work on referenda. Magleby challenged the idea that direct legislation improves voter participation. He analyzed voting on ballot propositions and concluded that voter turnout was neither increased nor made more representative by direct legislation. He found that ballot propositions were generally confusing, required

---

22 Responding to open-ended CV questions is analogous to participating in an oral auction or a first-price sealed bid auction. It is well documented that these generate lower prices than posted-price auctions [Plott (1989)].

23 Survey mode clearly affects response rates. For example, Infosino (1986) reports that, in AT&T's experience using telephone and mail to promote a new pricing option, three times as many customers exposed to the phone promotion purchased the product than those exposed to the mail promotion. A key factor with mail is whether people think the survey is junk mail and throw it out unopened. Seip and Strand (1992) made no allowance for the inherent non-response in mail surveys. Duffield and Patterson (1991) found a higher response rate with the University of Montana identified on the envelope as the sponsor than the Nature Conservancy. Apart from differences in response rates, the distributions of WTP responses were similar in their real and hypothetical payment experiments.

24 After the interview, one debriefs the interviewer to ascertain how attentive the respondent was, whether he seemed to understand the questions, whether he appeared confident in his responses, etc.

25 Most of the surveys in Hausman (1993) use open-ended questions, and none uses in-person interviews.
advanced reading skills, and were poorly understood by most voters.26 This actually strengthens the case for CV. Good CV studies go to great lengths to ensure that questionnaires are clearly worded, intelligible, informative, and impartial, and use probability samples to ensure a representative sample. Hence, a well-conducted CV study offers advantages over an election as a means of discovering people's preferences because one can control for bias in the information set and overcome the vagaries of voter turnout.27

DH also exploit another of Magleby's findings: public support changes over the course of an election. He found that 53% of the elections on propositions in California from 1960 to 1982 involved a reversal (i.e., the side that had once been losing ultimately won) compared to only 14% of the elections on candidates. He attributed this difference to the fact that, in propositions, voters were less knowledgeable about the issues, less sure of their voting intentions, and more susceptible to campaign advertising. Going a step further, DH claim that these reversals prove opinion polls inaccurately predict election outcomes and CV surveys are not an expression of individual preferences. Neither claim can withstand scrutiny.

Magleby himself identified advertising as a major cause of reversals on propositions. He found that negative advertising (unlike positive advertising) frequently determines ballot outcomes: opponents of a proposition "can virtually guarantee [its] defeat if they significantly outspend the proponents" (p. 147). This is illustrated by the case of "Big Green"— Proposition 128 on the 1990 California ballot—to which DH devote much attention. A poll in June showed 46% in favor and 38% opposed, with the rest undecided. In the November election, it lost by 36% to 64%. How did this happen? Opponents spent $11.9 million, while supporters spent $4.5 million. The positive advertising stressed the general need to protect the environment. The negative advertising stressed several points: Big Green would cause economic havoc; it was too complex and the regulatory provisions in the fine print could be damaging; and it was supported by Tom Hayden who would become pollution czar if it passed (the ads called it the "Hayden initiative").

26 In his book on referenda, Cronin (1989) had a very different view of voter rationality and rejected Magleby's position as exaggerated. Many others have seen a core of rationality in voter behavior, including Downs (1957), Fiorina (1981) and McKelvey and Ordeshook (1986) who emphasize how campaigners give out signals that overcome voters' limited knowledge. Lupia (1993) studied the insurance reform battle in the 1988 California ballot and concluded that the availability of informational "short cuts" enabled poorly informed voters to act as though they were well-informed. Snyder (1993) studied California ballots between 1974 and 1990 and found that "in the aggregate, preferences do not exhibit the instability that Converse and others have found at the individual level (so-called "non-attitudes")." Similarly, Page and Shapiro (1992) reviewed 50 years' of national opinion polls and concluded that "at the collective level, public opinion displays impressive characteristics of stability, rational responsiveness to concrete political situations, and the capacity to adapt as new issues become available and change the contours of these situations." DH do not mention any of the other literature on California referenda.

27 This has not escaped the attention of political scientists. Fishkin (1991) has proposed using "deliberative opinion polls," in which a random sample of the population is selected, given extensive information on an issue, and then polled—in effect, an extended CV survey. "Instead of a public opinion that is formed by inattention and the shrinking sound bites, it's a public opinion formed by debate, by thought, by face-to-face engagement" ["American in London to Test 'Deliberative' Polling" New York Times 9/21/93].
Does this prove that the California voters didn’t have stable preferences? Were the voters well-disposed to Tom Hayden in June, and ill-disposed in November? Surely not. It is more likely that, as the campaign unfolded, voters changed from seeing Proposition 128 as a single-attribute commodity to seeing it as a more complex, multi-attribute commodity. One attribute, evident from the beginning, was environmental protection. The opponents successfully added various other attributes unfavorable to the proposition. This provided voters with new information and forced them to make a trade-off among the attributes.²⁸

Opinion pollsters have long known that attitudes change during an election. For this reason, they avoid asking people about future intentions—they ask "if the election were held tomorrow, how would you vote?" rather than "how will you vote in November?" There is considerable evidence that polls reliably indicate public sentiment at the time they are taken, and that polls close to an election are generally accurate predictors of the outcome. DH cite an L.A. Times poll on Big Green to challenge the reliability of opinion polls. They claim the poll taken a week before the election found voters evenly divided. Their analysis is seriously incomplete. The Times article did state that Big Green trailed by only 2% among registered voters—44% were against, 42% for and 14% undecided, with a 3% margin of error. But, it also stated that, among voters considered the most likely to actually cast ballots, it trailed by roughly 12%. Assuming that most of the undecided would vote against the proposition, as is commonly done by political experts, the poll implied that voters would reject the proposition by a margin close to the actual outcome.

Similar results were found in a CV study of Proposition 25, the Clean Water Bond Law, just before the 1984 California election. This received little media attention and passed in November with 73% of the vote. Carson, Hanemann, and Mitchell (1987) purchased questions on the Field California Poll a month before the election to conduct a CV survey. Respondents were given a brief description of the proposition, told what it would cost their household, and asked how they intended to vote. The proposition was estimated to cost about $4/year per household. At this amount, 24% were undecided or did not intend to vote, 63% said they would vote yes, and 13% would vote no. If one assumed that between 50% and 70% of the undecided would ultimately vote no, this would have given a close prediction of the election outcome.

DH say that CV responses are more like opinion polls than a scientific evaluation of damages (p.30). This misses the point. The goal, as the NOAA Panel defined it, is to elicit people’s preferences as if they were voting in a referendum. The experience with real elections shows that advertising and turnout affect the outcome. CV deals with these by providing impartial information and using a random sample of the

²⁸ DH actually admit this but draw a strange conclusion: "Proposition 128 demonstrates that individuals do react to the level of information they have and the context of the situation in which they make decisions. Thus, the referendum claim for the use of CV has no foundation in individual economic preferences" (p. 17). Apparently DH believe that economic choices ought not to be based on information or context.

14
population. The other lesson from real elections is the need to be sensitive to "don’t know" responses and, generally, to treat them as votes against. This already is quite standard in CV studies. Beyond that, the available evidence suggests that the voting format needs little further calibration.

5. CV and Market Research

In market research, in contrast, calibration procedures are regularly applied to survey data on purchase intentions because of divergences between intended and actual purchases. Yet the NOAA Panel observed that these surveys must convey some useful information since firms continue to pay for them. But, more needs to be said. First, the degree of divergence, and therefore calibration, varies with circumstances, including the type of product and the investment in marketing effort. Urban and Hauser (1980) observe that "if the product is well-positioned and an aggressive marketing strategy is planned," a conservative estimate would be that 90% of those who say they definitely will buy the product actually do so. Second, as Juster (1964) pointed out, the reason why actual purchases are liable to differ from intended purchases is the intervention of unforeseen events and new information. This cuts both ways: some fraction of those who say they don’t intend to buy the item will end up buying it. If you discount the "yeses" but not the "nos," you typically underestimate total purchases.

However, this experience with market research might not apply to CV surveys. First, market research surveys usually provide considerably less information than a standard CV survey. For example, in a transportation study conducted at the University of Miami just before the opening of the Miami Metrorail (Sheskin (1991)), respondents were asked "Suppose Metrorail opens January 1, 1984; the price of gas remains as it is now; Metrorail costs $2.00 round trip, plus 25 cents for transfers to and from the bus; parking at a Metrorail station is $1.00/day. Would you use Metrorail to go to and from this campus?"

While this specifies some economic variables in admirable detail, it is incomplete in other respects. There is no mention of whether it would be hard to find a parking space at the station, whether the trains would run on time, the risk of getting mugged at night, or other attributes that could matter to users. Differences between the commodity described in the survey and the commodity subsequently experienced undoubtedly account for some of the divergence between intended and actual purchases. Second, people generally have control over the timing of their purchase of private goods, but not the timing of their voting on public goods. As an individual, you decide when your Ford gets replaced, but not your President. Some of the divergence between intended and actual purchases arises from changes in timing, which matters to market researchers in a way that doesn’t apply to CV. For example, in Juster’s study many respondents who said

---

29 Indeed, Mosteller et al. (1949) identified ignoring "don’t knows" as a major reason for the pollsters’ failure to predict the 1948 Presidential election, along with using quota samples and stopping their polls too soon.
that they would definitely buy an air conditioner or a car during the next six months, and then didn't, did buy the product at a later date, as opposed to losing interest and never buying it.

In addition, cognitive errors in consumers' perception of the duration of time may affect reports of purchase intentions, just as they affect consumer recall of past expenditures. There is abundant evidence of "telescoping" in the household expenditure survey literature, i.e., the tendency of respondents to allocate expenditures either to an earlier period than when they actually occurred (backward telescoping) or to a later period (forward telescoping) [Neter and Waksberg (1964)]. In the literature, forward telescoping has been found to be a more serious problem than backward telescoping, causing overstatement of past activity [Loftus et al. (1990)], and to increase in severity with the length of the recall period, e.g., recall of participation in sportfishing and hunting [Westat (1989)].

Telescoping is unlikely to be a factor in CV; but it certainly can affect travel cost analysis. An example occurs in Hausman, Leonard and McFadden (1993) [HLM], a travel cost study of the recreation losses from the Exxon Valdez oil spill. HLM used data from a 1990 telephone survey conducted for Exxon in which 8,888 Alaskans were asked how many trips they had taken in Alaska for sportfishing and other recreation during 1989 and 1988. The Alaska Department of Fish & Game (ADFG) has conducted a survey of sportfishing each fall since 1977, in which a fishing log is mailed to a random sample of fishing license holders. In 1989, it conducted a survey as usual and received responses from 14,517 anglers [Mills (1990)]. Here, then, are two surveys intended to measure the same thing—one done for litigation, the other done to generate economic data for fisheries management. The Exxon survey, conducted in the spring of 1990, produced an estimate of 2,464,000 fishing trips by Alaskan residents in 1989 and 2,480,000 trips in 1988. The ADFG survey, conducted in the fall of 1989, covered both resident Alaskan and non-resident anglers and produced an estimate of 1,731,000 fishing trips in 1989, 30% less than estimated from the Exxon survey. This result is consistent with the thesis that a longer recall period generates more forward telescoping. The 1988 ADFG survey showed 1,919,000 trips in 1988, 23% less than estimated from the Exxon survey. This significantly affects the interpretation of the two surveys. HLM intended to use their 1990 survey of fishing activity in 1988 as a control for the effects of the oil spill. Based on this control, HLM found that the spill caused a decline of 16,000 sportfishing trips, while the ADFG data showed a decline of 188,000 trips. In relative terms, this is a decline of 0.65% versus 9.8%.

With CV, DH apply the criterion that "because two ways measure the same quantity, they should yield approximately the same answer" (p.47). This is clearly violated with the two sportfishing surveys. Should all recreation surveys therefore be discontinued? Should we give up on the travel cost method? Surely not.

---

30 HLM's only reference to the ADFG data is a footnote saying that it "finds a larger decline in number of trips than does our survey" (p.360). This is a masterpiece of understatement.
We should, however, recognize that surveys are fallible and that careful design and implementation are just as necessary for surveys used in revealed preference as in CV. Admitting these limitations should not cause us to abandon the effort in either case.

6. CV Studies Commissioned by Exxon

In the summer of 1989 Exxon retained a distinguished group of economists to investigate the CV method. One might have thought their empirical work would focus on the Exxon Valdez oil spill and CV’s ability to measure its damages. Not so. The principal surveys reported in Hausman (1993) deal with other issues—preventing logging in wilderness areas in the Rockies, covering waste-oil holding ponds in Texas, and improving the capability for handling small oil spills that occur routinely each year in U.S. coastal waters. These items could have been studied at any time before 1991. But, the surveys were all initiated in March 1991, just after plans were announced to settle state and federal suits against Exxon, and at a time when two federal agencies were gearing up to issue new regulations for natural resource damage assessment. The surveys went into the field in June. The entire process of instrument development, testing, and field preparation were compressed into the period March - June, 1991.²¹

DH justify their conclusion that responses to CV surveys are not an expression of underlying preferences by demonstrating inconsistencies between the responses to these surveys and what they claim are predictions of economic theory. This begs many questions. How do DH know the respondents’ true preferences? Could the responses be affected by ambiguities or flaws in the surveys? Is DH’s interpretation of economic theory correct? These possibilities must all be eliminated before one may conclude that CV is a defective measurement tool. In this section I focus on issues concerning the survey design and data analysis. In the next section I take up the arguments about economic theory.

One group of surveys dealt with WTP to protect some of 57 federal wilderness areas established in Colorado, Idaho, Montana, and Wyoming under the 1964 Wilderness Act from being leased for commercial development to reduce the federal deficit. The survey was read over the phone to a random sample of households in the four states. Respondents were told that the development would involve harvesting mature timber at a rate of 1% per year, which would require building roads and bringing in heavy equipment. Of the 57 areas, 7 had already been proposed for leasing. Respondents were asked their WTP to preserve another area that might be considered for a similar timber harvest, the Selway-Bitterroot Wilderness (1.3 million acres in northern Idaho). Other versions of the survey proposed logging in two

---
²¹ Despite promises made at the Exxon Conference in April 1992 and repeated in Hausman (1993), the raw data from these surveys still have not been made available to outside researchers. The published papers provide summary statistics on WTP responses. In some cases, it has been possible to apply techniques for grouped data [Haitovsky (1973)] to reconstruct what the regression equations must have been, see Carson and Flores (1993).
other wilderness areas, individually or in combination: the Bob Marshall Wilderness (1 million acres in Montana), and the Washakie Wilderness (700,000 acres in Wyoming).

The lack of detail in this survey instrument is striking. Standard practice would be to list the wilderness areas by name and provide a map. The survey should at least name the seven areas already proposed for leasing. The questionnaire doesn’t explain how the named wilderness area differs from these other areas, nor why it is being singled out for attention. Nor does it say what would happen to these seven areas if the named area were preserved, or what would become of the 49 other wilderness areas—if the Wilderness Act can be set aside, don’t these areas need protection, too? Failure to specify such details is harmful in two respects. First, the quantity of wilderness protection that respondents are "buying" is ill-defined. Since people’s default assumptions about this may vary, respondents may be valuing different commodities. DH lost control of their valuation exercise at the very start. Second, the lack of detail creates an air of unreality which may deflect respondents from taking the survey seriously.

This is exacerbated by incongruities between the scenario and events in the area’s recent history. When the areas featured in the survey were first designated there had been considerable controversy over the possibility of mining development. The 1964 Wilderness Act permitted new mining claims in designated wilderness areas through 1983. There was a public outcry in 1981-2 over applications to lease two-thirds of the wilderness acreage in Washakie for oil and gas exploration, with potentially serious adverse impacts on elk, moose, bighorn sheep and grizzly bears. This was opposed by the Wyoming Congressional delegation, not normally known as tree-huggers, and was quashed. Respondents might have taken the survey’s mention of "commercial development" to refer to something like this. Also, there was some public discussion of wilderness issues in 1991 due to an approaching deadline for adding areas to the National Wilderness System. But, this debate concerned proposals to designate new areas as wilderness, not to remove existing designated areas. For the scenario in the survey to be feasible, it would have been necessary to revoke the Wilderness Act, which would have been front-page news in the region.

Diamond et al. use these survey responses to test four sets of hypotheses which, they claim, expose incongruities between the CV responses and economic theory—hypotheses about the effects of size on WTP, substitution effects, embedding, and the additivity of WTP. The data do not appear to support their claims. Diamond et al. test the null hypothesis that WTP does not vary with size for the three featured

32 Without a map, names can be a problem in this area. For example, many guidebooks don’t list the Washakie Wilderness by that name; and there are several Roadless Areas that are called Selway-Bitterroot. Although DH don’t cite it, there was an earlier CV study of the Washakie Wilderness and it did use a map [Barrick and Beazley (1990)].

33 A colleague has remarked that this is like testing the accuracy of public opinion polls by conducting surveys in which people are asked: "If the Presidential election were being held today and Bill Clinton were running against a tall guy from Maine and a short guy from Texas, who would you vote for?" and "If the Presidential election were being held today and Bill Clinton were running against two tall guys from Maine ...." Why would anyone do this?
wilderness areas and find that it cannot be rejected. However, they use a nonparametric test which has very low power in this case [Carson and Flores (1993)]. A more natural approach is to regress WTP on size. When this is done, the null hypothesis is strongly rejected. The test for the absence of substitution effects takes the form of comparing the WTP for a featured site when respondents are told that seven other unnamed sites have been opened to logging versus eight other sites. Diamond et al. argue that the WTPs should differ if substitution effects are present. They test the null hypothesis of no difference in WTP for Washakie and Bob Marshall, again using nonparametric tests of low power, and find it cannot be rejected. For Washakie, however, if one employs a simple t-test instead of Diamond et al.’s nonparametric test, the null hypothesis is rejected and their conclusion is reversed.

Their embedding test focuses on whether WTP increases as one switches from a less to a more inclusive commodity. They approach this by testing a single null hypothesis that WTP is the same across surveys involving one, two and three featured wilderness areas, and find it cannot be rejected. However, if one compares WTP for individual sites taken alone versus in larger groups, the null hypothesis is rejected in about half the cases, as it is if one compares WTP for one site versus for all sites. They test for additivity—e.g., whether WTP to save both Washakie and Bob Marshall equals the sum of the WTPs to save each site taken separately—and find it can be rejected in favor of sub-additivity. However, as noted earlier, the sub-additivity of WTP is simply Madden’s (1991) property of R-substitutability and should have come as no surprise.

McFadden and Leonard (1993) [ML] used data from other versions of the wilderness survey, including versions with a closed-ended WTP question. There are several useful innovations including a Box-Cox formulation for the utility function generalizing the linear and logarithmic models in Hanemann (1984). ML raise many of the same issues as Diamond et al. (e.g., repeating their erroneous claim that sub-additivity is “inconsistent with classical preferences”), while adding new ones concerning the effects of elicitation format and the income elasticity of WTP. ML make much of their finding that open-ended questions yield lower WTP estimates than closed-ended questions (p.166). However, this has been documented in the CV literature for a decade. It is explained by the factors I listed in Section 3, and is not per se inconsistent with economic theory. Likewise ML’s finding that single- and double-bounded versions of the closed-ended format produce different results has already been noticed and discussed [Hanemann, Loomis and Kanninen (1991)]. In Hanemann (1991b), I showed further that there can be an inconsistency between the first and second responses in the double-bounded format. But, the evidence is that the double-bounded format generates lower WTP values and this difference arises from an increased

---

tendency to say "no" to the second bid rather than "yes." Respondents having been told that the item can be provided at one price, now resist when told that the price might be higher. In statistical terms, the double-bounded format provides more information for a given sample size and greatly shrinks the confidence interval; in using it, one is trading off a large reduction in variance for some downward bias.

ML make much of the income elasticity of WTP for wilderness. They equate this with the Box-Cox parameter $\lambda$, which they estimate at 0.269. They consider this value is so low that it casts doubt on CV’s validity (pp. 185-6). Yet ML have erred in calculating the income elasticity. It is not the same as $\lambda$ and, indeed, varies inversely with $\lambda$. The correct estimate in their case is 0.731. This aside, it is hard to see how ML can make any statement about the correct value for an elasticity of WTP without knowing people’s true preferences. Moreover, income elasticities of demand for state and local government services in the existing literature are generally well below unity; they mostly range from 0.3 to 0.6 [Cutler, Elmendorf and Zeckhauser (1993)].

Another survey conducted for Exxon has received considerable attention, the "bird" survey by Research Triangle Institute (RTI). This survey provides considerably more detail than the wilderness survey. It tells respondents that migratory waterfowl spend the summer in the northern parts of North America and fly south for the winter using four flyways, as shown in a map. The largest is the Central Flyway, used by 8.5 million waterfowl. Along the route, in parts of Texas, Oklahoma and New Mexico, there are over 250,000 waste-oil holding ponds ranging in diameter from 10 to 100 feet. Migratory waterfowl are attracted to these ponds because there are so few wetlands along the flyway and drown when they land on the ponds. To prevent this, the federal government is considering requiring owners to cover the ponds with fine wire mesh nets that would keep waterfowl safe. If the regulation is approved, consumers will face higher prices for petroleum products. Respondents are then asked their annual household WTP for this program in an open-ended question. There are 3 different versions of the survey. The first version says that about 2,000 migratory waterfowl die in the ponds; the second says about 20,000; and the third about 200,000. The program would prevent these birds’ death. The 3 versions were implemented as self-administered interviews with three separate samples of shoppers at two malls in the Atlanta area, averaging 10-12 minutes per interview.

While closer to standard CV practice than the wilderness surveys, the bird survey has some critical flaws. For example, Desvousges et al. (1992) mention that some people in the pre-tests worried that the waste-oil ponds might be harmful to children; therefore they added to the text that the ponds were located in "remote and sparsely populated areas". But, no evidence is presented that this wording eliminated

---

35 Denoting the income elasticity of median WTP by $\eta$, to a first approximation one has $\eta = 1-\lambda$. The only way to get an income elasticity greater than 1 in the Box-Cox model is to have a negative $\lambda$: when $\lambda > 1, \eta < 0$. This is an undesirable feature of the model. ML do not mention Kristrom's (1990) alternative flexible utility model.
human safety as a perceived benefit from the program. There is also the credibility issue: did respondents believe that, with 8.5 million birds flying over one quarter of a million ponds, and with a federal program in the works, there could be only 2,000 bird deaths per year?

What is the real variation in scale? In terms of the absolute number of bird deaths, the three versions of the questionnaire differ by two orders of magnitude. But, the contrast perceived by respondents to the three versions may have been much smaller. Desvousges et al. (1992, p.36) note that people often want injuries expressed as percentages—they want to know the relative rather than absolute impact on a population. This information was incorporated in the survey. In the 2,000 bird version, the text said this was "much less than 1 percent" of the waterfowl in the Central Flyway; in the 20,000 bird version, it said "less than 1 percent"; and, in the 200,000 bird version, it said "about 2 percent." If respondents focused on the relative impact, it is hard to believe that they would have perceived any significant difference in this tiny range of percentages. Moreover, there were no other cues, visual or verbal, to highlight the contrast among the three versions, apart from changing a few words in a text running to a dozen pages.

The lack of contrast was almost certainly exacerbated by the survey mode—intercepting shoppers as they walk around a mall, promising them a fee if they participated, and having them sit down with their bags for ten minutes to fill out a self-administered questionnaire. In a well-designed CV study, one wants respondents to reflect on the issue at hand and give a considered opinion. Unless the study deals with a consumer product, shopping mall intercepts are a poor way to accomplish this.\footnote{In addition to distracted shoppers, another hazard of mall intercepts is infiltration by teenagers. RTI's protocol for the bird surveys called for screening to exclude participants under 20. However, the lowest age group (20-29), where "illegal" teenage participants would have been assigned if they had penetrated the study, accounts for 46% of the sample and for 73% of the respondents subsequently identified as outliers because they listed unusually high incomes and gave WTP responses of $1,000 or more.}

All of this casts doubt on the ability of the bird survey to provide a meaningful test of the responsiveness of WTP to variation in the scale of the item being valued. This is compounded by statistical problems arising from the survey format and mode. It is an object lesson in why one should avoid an open-ended format: 17% of respondents could not give a numerical answer to the WTP question compared, for example, to 2% with a closed-ended format in another CV study by RTI. Of the rest, about 6.5% gave WTP amounts of $1,000 or more, and 1.8% gave amounts over $12,000 or that exceeded a quarter of their annual income. Thus, the data are noisy and show signs of gross contamination unusual for a high quality CV survey.\footnote{Payne and Schkade (1993) provide evidence of a survey mode effect, since they use the same instrument but administer it differently. In their study, respondents recruited by phone came to a central location, responded to the questionnaire individually in the presence of an interviewer, and were asked to think aloud as they answered. Compared to RTI's self-administered mall intercepts, this procedure slowed respondents down and forced them to think. While RTI's respondents took about 10 minutes to complete the questionnaire, these respondents took about}

\footnote{Payne and Schkade (1993) provide evidence of a survey mode effect, since they use the same instrument but administer it differently. In their study, respondents recruited by phone came to a central location, responded to the questionnaire individually in the presence of an interviewer, and were asked to think aloud as they answered. Compared to RTI's self-administered mall intercepts, this procedure slowed respondents down and forced them to think. While RTI's respondents took about 10 minutes to complete the questionnaire, these respondents took about}
Desvousges et al. (1992) provide three treatments. The treatment reported in Hausman (1993) removes about three quarters of the responses of $1,000 or more, leaving the rest; while the mean WTP does differ among the three bird scenarios, the differences are not statistically significant. The second treatment trims 5% of the responses from each tail, which is roughly equivalent to the first treatment and yields similar results. The third treatment trims 10% from each tail. Then, differences in mean WTP are statistically significant: Carson and Flores (1993) find that, if one regresses WTP on the percent of birds killed, the slope coefficient is significant at the 95% level.\footnote{If the WTP distributions for the three scenarios were the same, as Desvousges et al. maintain, trimming would have no effect on the outcome of tests for the equality of moments, as happens here. Hampel et al. (1986 p. 401-2) conclude that, in the presence of gross contamination, applying procedures like trimming which are usually associated with symmetric distributions to asymmetrically distributed data is much preferable to making no adjustment at all.}

To summarize, the Exxon CV surveys have some serious flaws. To regard them as representative of good practice in CV or a crucial test of the CV method makes sense only if one believes that measurement results are invariant with respect to measurement practice. This is not true in economics or in any other science.

7. CV and Economic Theory

DH's argument that CV is incompatible with economic theory rests on the claim that CV responses are inconsistent with what economic theory says about people's preferences, as well as the fact that CV violates economists' inclination towards revealed preference. Both raise issues that are worth discussing even apart from the debate on CV.

As we noted above, DH's assertion that CV responses fail to vary with the scope of the commodity is not supported by Exxon's CV studies. Nor is it supported by the empirical literature on CV, to which DH make no reference. In one of the earliest CV studies, Cicchetti and Smith (1973) looked at how WTP varied with commodity characteristics. They asked hikers in a Montana wilderness area about their WTP for trips involving different levels of solitude, as measured by the number of encounters with other hikers on the trail during the day or in camp at night. These trip attributes had a statistically significant effect on WTP. For trips where other hikers were encountered on two nights the WTP was 34% lower than for trips with no such encounters. Many other studies have since reported comparable findings using both internal and external tests of scope.\footnote{Kahneman's (1986) CV study is widely cited as showing that people are willing to pay the same amount to clean up fishing lakes in one region of Ontario as in all of Ontario. But, his graph actually shows a 50% difference in median WTP. For evidence of significant scope effects see the meta-analyses in Walsh, Johnson and McKean...}
How much should WTP vary with scope? Diamond (1993) asserts that the elasticity of WTP with respect to scope should exceed unity—people should value preventing 200,000 bird deaths over a 100 times more than they value preventing 2,000 bird deaths. To prove this, he starts off with a model where people care about the number of birds originally in the population, \( q_0 \); the number at risk of dying, \( q_R \); and the number of those that are saved, \( q_s \). Letting \( z \) denote a Hicksian composite market good, the utility function \( u(q_0, q_R, q_s, z) \) is decreasing in \( q_R \) and increasing in the other arguments. If \( u \) were quasiconvex in \( q_R \), the elasticity of WTP would exceed one. Diamond imposes this restriction indirectly. He first assumes that \( u \) is quasiconcave in \( q_0 \) and then imposes the restriction that \( q_0, q_R, \) and \( q_s \) are perfect substitutes, so that \( u(q_0, q_R, q_s, z) = w(q_0 - q_R + q_s, z) \) for some function \( w \). If \( u \) is to be quasiconcave in its first argument, which then makes \( w \) quasiconvex in \( q_R \).

Arrow (1986) and Simon (1986) warned against the tendency in contemporary economics to make assertions about rational behavior that depend primarily on auxiliary, factual assumptions. As Simon put it: "Almost all the action, all the ability to reach nontrivial conclusions, comes from the factual assumptions and very little from the assumptions of optimization. Hence it becomes critically important to submit the factual assumptions to careful empirical test." Diamond's claim about the elasticity of WTP provides an excellent illustration. It follows not from rationality per se but, rather, from the assumption of perfect substitution, about which economic theory is silent. Let \( q_i = q_0 - q_R + q_s \). Diamond asserts that \( u(q_0, q_R, q_s, z) = w(q_i, z) \)—people should care only about the ultimate population size, not how many were in the original population, at risk, or saved. When CV data disconfirm this assumption, Diamond dismisses the method. Others might be more inclined to believe the data and reject the assumption.

Diamond's argument violates a foundational premise of economics. As Simon noted: "Neoclassical economics provides no theoretical basis for specifying the shape and content of the utility function." Hausman (1993), too, abounds with prescriptions that violate this precept. People should care about outcomes, not about the process whereby they are generated; they should not care whether animals are killed by humans or die naturally; they should care only about the number of acres of wilderness; they should not care whether development takes the form of logging or mining; they should not be concerned about small or reversible injuries; they should value things only for selfish motives. With this litany of "economic correctness," consumer's sovereignty is shoved aside.\(^{40}\)

For welfare economics, DH adopt an extreme position. They follow Milgrom (1993) in contending...
that behavior cannot be considered an acceptable expression of preference if it is motivated by ethical concerns. For nonuse value to be validly included in benefit-cost analysis "it would be necessary for people's individual existence values to reflect only their own personal economic motives and not altruistic motives, or sense of duty, or moral obligation" (p. 431). This is hardly consistent with the general practice in economics. We did not remove Catholics when estimating demand functions for fish prior to Vatican II. Nor do we adjust when childless couples vote for school bonds. The modern theory of social choice considers it immaterial whether preferences reflect selfish interest or moral judgment: "It need not be assumed here that an individual's attitude toward different social states is determined exclusively by the commodity bundles which accrue to his lot under each. The individual may order all social states by whatever standards he deems relevant" [Arrow (1960 p.17)].

DH are also at odds with the modern approach to demand analysis, due to Lancaster (1966). He rejected the "puritanical view" that consumers should care just for the raw quantity of commodities and proposed instead a model in which they care for characteristics and how these are combined to form a consumption experience. It is characteristics, not commodities, that give rise to utility. Any commodity possesses multiple characteristics. For example, "a meal (treated as a single good) possesses nutritional characteristics but it also possesses aesthetic characteristics, and different meals will possess these characteristics in different relative proportions" (p. 133). This makes demand context dependent. Lancaster also noted that the units in which characteristics are measured may be arbitrary. They may be ordinal or cardinal. He emphasized that, while the characteristics of a commodity are an objective fact, people's reaction to them is subjective; e.g., some may find a shape beautiful, others not. Being subjective, the preference for characteristics is likely to vary among individuals: "some [characteristics] may be relevant to one individual, others to a different individual. One person may scarcely notice the existence of properties which to him are of little account in his decisions but are important to someone else." For this reason, aggregate behavior cannot be expressed in terms of the traditional representative consumer. Above all, Lancaster stressed, "goods are simply what consumers would like more of; and we must be neutral with respect to differences in consumer tastes" (p. 132). The same conclusion surely holds for CV, which applies Lancaster's model to environmental commodities specified in terms of characteristics.

41 Milgrom also claims that using CV to measure altruistic preferences creates double counting. His case is not convincing. First, it depends on the particular specification of the utility function, as Johansson (1992) notes; if the argument of the utility function is another's consumption rather than his utility, there is no double-counting. Second, it derives its force from the auxiliary assumption that the respondent does not realize that the other people for whom he cares will have to pay, too; this is not a problem in a referendum format. Third, in many CV studies the object of the altruism is often wildlife—sea otters, for example. Since the wildlife are not surveyed, the issue is moot.

42 Cf. Adam Smith (1776): "The whole industry of human life is employed not in procuring the supply of our three humble necessities ... but in procuring the conveniences of it according to the nicety and delicacy of our taste."
Lancaster saw traditional demand theory as a "coarse structure" theory appropriate to broad aggregates of goods such as "automobiles," "food," and "clothing" rather than individual goods as strictly defined. There is necessarily less substitution among broad commodity aggregates than among the individual goods within a group. This should be borne in mind when one considers the embedding problem in CV. As noted above, this is a direct consequence of Madden's (1991) R-substitution. The greater the degree of substitution, the more extensive the sequencing effect that economic theory predicts for any one of a set of commodities. Substitution elasticities estimated for traditional commodity aggregates will understate the substitutability, and thus the sequencing effects, to be expected among individual commodities.

DH's ultimate argument against CV is that it violates economists' traditional commitment to revealed preference. Revealed preference has undoubtedly served economics well. But, this does not mean that it lacks shortcomings. As a profession we tend to be blind to them. Revealed preference is a reflection of the behaviorist movement that dominated Anglo-American psychology from the early 1920s to the mid-1950s. It was introduced into economics in 1938 by Paul Samuelson in his first published paper, which sought to eliminate from consumer demand theory "any vestigial trace of the utility concept." Hutchison (1938), published the same year, introduced positivism and Popper to English-speaking economists. These two works were highly influential in establishing the position that inference and analysis in economics should be based solely on external observation of behavior.

As both a logical principle and a methodological approach, behaviorism suffers from some severe problems that were exposed in the 1950s by the cognitive movement in psychology (Baars (1977)) and the linguistic movement in philosophy (Goldman (1993)). Much of the discussion in the debate centered on logical objections. But, in the view of my colleague John Searle, a leading authority on the philosophy of mind and language, "it is the commonsense objections that are the most embarrassing. The absurdity of behaviorism lies in the fact that it denies the existence of any inner mental states in addition to external behavior. And this, we know, runs dead counter to our ordinary experience of what it is like to be a human being" (Searle (1992 p.5)).

Forty years after it waned in the other social sciences, behaviorism continues to hold sway in economics. But, there has always been criticism, starting with Frank Knight's (1940) review of Hutchison. The subject matter of economics, Knight wrote, is human conduct and motivation, not the study of mechanical response. Testing predictions is ultimately a social phenomenon. The fundamental propositions of economics are not inferred or verified in the same way as those of mathematics. The main source of

---

43 Samuelson's subsequent retreat from this radical position is analyzed by Wong (1978).

44 For this reason, Searle notes, behaviorists were accused of "feigning anesthesia" as in the story of the two behaviorists. After making love, one turns to the other and says "Well, it was great for you. But, how was it for me?"
knowledge about human conduct is interaction with humans. Economic behavior involves intentions, which are not amenable to external observation; the imperfect link between motive and result is proof that one does not infer the former from the latter. "The difference between predicting human behavior and predicting the behavior of physical objects ... [is] that the latter neither behave irrationally or sentimentally, nor make mistakes, nor change their minds... [This] is admittedly embarrassing to the economist as a scientist, but there does not seem to be anything that he can do about it" (p. 29). 45

The problem with revealed preference, as Sen (1973) and others have noted, is that like all induction it requires an act of faith to extrapolate from particular choices to general assertions about behavior and preference. One needs a host of auxiliary assumptions to rule out factors that might invalidate the extrapolation. Choice may have little to do with individual preference, perhaps because preferences are incomplete, or choice is a social not an individual act. Tastes may vary among individuals; tastes may change if sufficient time elapses. Since revealed preference compares the behavior of different individuals, or the same individual at different times, in response to different prices, how can we tell what part of the difference in behavior is due to price and what part to a difference in preference? The *ceteris paribus* condition may be violated in other ways. For the Weak Axiom to hold over all market prices, one must observe the individual’s choices under infinitely many price configurations. Such problems must be assumed away in order for revealed preference to work. But, the assumptions are not verifiable if one is restricted to observed behavior. This makes revealed preference a somewhat hypothetical undertaking. 46

One of the key assumptions of revealed preference in practical applications is a "universal" choice set -- i.e., the consumer chooses among the complete set of all possible alternatives and, in choosing any one item, has full regard for all his other consumption decisions. The reality is otherwise. After my own travel cost studies of beach recreation in the Boston area and sportfishing in Alaska, I concluded that few if any individuals were consciously choosing from all the sites that I knew to exist. Different individuals were choosing from different choice sets. One ought to model behavior as a two-stage process, first the formation of a choice set and then a conventional maximizing choice within that set. In addition, many people’s recreation choices were separate from their other consumption decisions, in the manner not of a separable utility function but rather of piecemeal decision making: when going fishing, think of recreation alternatives; when buying a sofa, think of furniture alternatives; don’t think of sofas when going

45 Knight noted the irony that, having waged a long struggle to escape from the idea that stones are like men, we now seem intent on showing that men are like stones.

46 For welfare evaluation, the *cut-off price* at which demand for the commodity ceases may introduce another hypothetical element into revealed preference. Though necessary to estimate consumer’s surplus, this may be outside the range of the observed data (e.g., one has travel cost data only for participants, or one believes that participants and non-participants have different preferences). A case where, for this reason, travel cost yields a less reliable estimate of WTP than CV is Hanemann, Chapman and Kanninen (1993)
fishing or vice versa. This form of satisficing where the choice set is edited so as to tailor it to the
decision at hand is not uncommon, but it is overlooked in consumer theory. We similarly assume a
"universal" attribute set: the same q's figure in everyone's utility function. This too is false. The q's differ
among individuals and over choices. The earlier discussion of Big Green illustrates this -- what initially
was a question of protecting the environment later became a question of Tom Hayden's political
standing. In short, the Achilles heal of revealed preference is that you have to know what the choice
is about. Without asking, you may not be able to tell.

These problems with revealed preference are counterparts to the problems raised about CV, such as
framing effects and embedding effects. The problems are not widely discussed in the demand literature
because they are kept from sight through the maintained assumptions about the choice set, the attribute
set, and the homogeneity of preferences across individuals and choice situations. With individual CV
responses, they are more readily apparent. But, these are general phenomena of human behavior, not
artifacts of the CV method. As Sen observes, "we have been too prone, on the one hand, to overstate the
difficulties of introspection and communication and, on the other, to underestimate the problems of
studying preferences revealed by observed behavior" (1973, p.258).

8. Conclusions

It certainly is true that no number can sometimes be better than some number. Is it true that no
number must always be better than some number? Is it true, as DH assert (p. 29), that CV surveys never
measure people's preferences and are never a suitable source of information on values in either benefit-
cost analysis or damage assessment? The NOAA Panel saw these as these "extreme arguments," and
rejected them [Arrow et al. (1993, p.41)].

DH take the position that ordinary people are hopelessly ill-informed, liable to exaggerate the damages
from oil spills, and should not be trusted to make judgements concerning monetary values. This should
be decided by experts, by "people like you or me," as Hausman told the NOAA Panel [NOAA (1992,
p.52)]. Why draw the line at oil spills --why not let economist philosopher-kings decide the value of all
public goods? This raises an ethical question of whether, as a society, we should leave important public
decisions for an intellectual elite to decide. It also raises the logical question of how the experts
themselves know the monetary value of the damages.

Empirical demand analysis traditionally dealt with a representative consumer and broadly defined
commodities. This was driven by the type of data available, which were generally highly aggregated over

47 The variation in the q's is clearly recognized in political science. Riker (1990), for example, distinguishes two
forms of political persuasion: rhetoric, which he defines as changing people's opinions about issues; and what he
calls heresthetic, which is changing their conception of what issues are at stake.
both individuals and commodities. Yet our observation of economic life shows that this is too simplistic. Moreover, over the last 30 years economic theory has developed a much richer view of individual behavior in both market and non-market settings. This includes not only Lancaster's work on consumer demand but also all the work on the economics of the family since Becker and on the economics of public choice since Downs. CV provides one way in which the more realistic theory of individual behavior can be given practical expression. It provides one way of tracing out the demand curve for a commodity that cannot be revealed through market data, but nevertheless exists and should count in an economic analysis.


McMillan, Paul, "Is Sympathy an Economic Value? Philosophy, Economics, and the Contingent Valuation


