

BEHAVIORAL FOUNDATIONS OF ENVIRONMENTAL ECONOMICS AND VALUATION

[John K. Horowitz](#) and [Kenneth E. McConnell](#)

Department of Agricultural and Resource Economics
University of Maryland
College Park MD 20742

[James J. Murphy](#)

Department of Economics
College of Business and Public Policy
University of Alaska
Anchorage, Alaska 99508

August 1, 2008

1. Introduction

For at least 60 years, economists have worked on empirical approaches to measuring the value of non-market goods and services. In its beginnings, this research employed models of revealed preferences, such as the travel cost approach for recreation¹ or the hedonic property value model for air pollution.² Economists pursued valuation based on revealed preferences because they were initially interested in what now, from our vantage, appears to be a narrow range of non-market services—those services for which revealed preference approaches could be applied. Further, given the profession's early suspicion of attitudinal surveys³, efforts to use any sort of stated preference approach would have met with even more intense opposition than they faced decades later.

As economists became interested in the valuation of a wider range of non-market services the shortcomings of revealed preferences became more apparent. Revealed preference techniques could be used as long as the behavior of interest led to the correct welfare measure and was observable, but in numerous circumstances observed behavior was not available. One could not have employed a revealed preference study of the value of reducing pollution in Lake Erie in the 1970's because the lake was so polluted that there was little use and no alternative, comparable, cleaner lake to observe. In this case, there is no revealed preference data on which to base valuation, and hence no ability to estimate the value of pollution reduction. In other

situations, environmental quality was not sufficiently connected to specific, identifiable behaviors. Before suspended particulates were suspected of health effects, it was clear that the pollutant soiled many buildings, a damage with no obvious behavioral implications (see Bockstael and McConnell, pp. 50-51). The absence of data or the minor and almost unobservable changes in behavior made revealed preference models impractical. Health effects such as coughs and sore throats were associated with averting behavior that was individually rather minor and hence could not be reliably observed, even though the aggregate values were large. Hence stated preference approaches to valuation were developed out of the necessity of doing valuation in the many circumstances where revealed preference methods might have been appropriate but would not work. Early efforts to value visibility at national parks (Rowe and Chestnut) illustrate the necessity of stated choice methods for valuation.

The failure of revealed preference methods for valuation tasks was the first impetus for developing stated preference methods. The emergence of the notion of non-use values provided a second and perhaps more compelling motive for developing stated preference approaches. Beginning in 1967 with Krutilla, economists developed the idea of non-use value, also known as passive use or existence value, a component of valuation that explicitly was not associated with observed behavior.⁴ Krutilla stated the idea clearly, noting for example that the existence of a fragile ecosystem is part of the real income of many individuals but does not contribute to the area under the demand curve for the resource. This came to be called existence value later. Pure public goods with substantial existence values such as visibility, regional air quality or pristine environments could not be valued with revealed preference approaches but were important for environmental policy. There is no better example than the damages from the Exxon Valdez oil spill.

Stated preference studies now make up a large proportion of valuation research. This is not simply for the original reasons—the inability to observe some actions and the need to measure existence values—but the growing recognition that econometric problems compromise many preference studies. Empirical studies now face the presumption that estimated models will be contaminated by unobserved individual heterogeneity unless proven otherwise.

In this chapter we are concerned with two problems that have arisen as economists have applied stated preference approaches to valuation. In particular, we review two issues – differences between values derived from real and hypothetical surveys and the gap between

willingness to pay and willingness to accept – that are crucial to the acceptance and advancement of stated preference techniques. The NOAA Blue Ribbon Panel identified both of these issues as problems for the use of contingent valuation in damage assessment. The Panel recognized from the extant empirical evidence that subjects in CV studies were likely to overstate their real willingness to pay. While the Panel acknowledged the minimum compensation that would be accepted would be the correct measure in many cases, they recommended estimating only willingness to pay because respondents would ‘give unreasonably high answers’ to willingness to accept questions.

We focus on real versus hypothetical valuation surveys because real surveys are presumed to be closer to ‘true value’ than hypothetical surveys. By this reasoning, a necessary condition for valuation surveys, the vast majority of which are hypothetical, is that they use features that can be shown to approximate real value surveys in experimental settings. Economists believe that valuation exercises should approximate ‘true value’ because these values are used in benefit-cost analysis and damage assessment, where real money is at stake.

We focus on the WTA/WTP disparity because it presents a challenge to valuation even when hypothetical surveys can be shown to approximate real surveys. High WTA/WTP ratios are perhaps the best and most widely documented contradiction to the neoclassical welfare model that underlies benefit-cost analysis. Furthermore, even if one accepts a role for valuation in the absence of a neoclassical framework, the WTA/WTP disparity shows that surveys can be highly sensitive to survey features that are not part of the valuation framework; in this case, the initial assignment of property rights.

The WTA/WTP disparity is an example of a broader phenomenon in which survey responses are unexpectedly sensitive to features that would seem to be incidental to environmental valuation. The counterpart is the set of phenomena in which survey responses are relatively insensitive to features that would seem to be important to environmental valuation. We review these issues briefly in Section 4. For reliable use of contingent valuation with neoclassical demand theory, it is necessary to rationalize the gap. The alternative is to employ stated preference methods but use the appropriate measure that depends on property rights, as Knetsch (1995) has argued. Nevertheless, because so much of the evidence on the *WTP-WTA* gap relies on stated preferences, making sense of the gaps is an essential component of sustaining the validity of this valuation method.

A central theme of this chapter is that valuation in the form of stated preferences and experiments are unavoidably intertwined. Holt and Davis described four forms of economic experiments. This current volume organizes these into market institutions, social dilemmas, voting/coordination games, and valuation. Of these four, only for valuation can experiments be said to be indispensable, essential to the conceptualization and understanding of the underlying economic ideas. Each of the other topics has real-world counterparts that truly interest us; that is, when experiments fail to replicate the real world, the experiments can be put aside; the real world trumps all. For valuation, however, there is no real world ‘check.’ All of what we know about valuation we learn from experiments.

Valuation is a form of experimentation and this experimentation has played a large role in learning about preferences and by extension, behavioral economics. In this chapter we will assess the development and recent findings concerning what we see as the two major challenges to valuation, namely the disparity between willingness-to-accept and willingness-to-pay and the connection between hypothetical and real valuation exercises.

Willingness to accept and willingness to pay

The welfare measures WTA and WTP are intuitively suited for policy analysis and fit neatly in neoclassical demand theory. Willingness to accept is the amount of money that would make an individual indifferent between a particular bundle of goods and a different, lesser bundle of goods plus the extra money. Willingness to pay is the amount of money that would make an individual indifferent between the lesser bundle of goods and the greater bundle of goods but absent that amount of money.

The experimental approach

The experimental approach is an essential component of valuation. Almost all of what we know about valuation is based on experimentation—in the lab and in the field. We learn the existence and sources of the WTA-WTP gap through experimentation. Carefully designed experiments can include both hypothetical and real payment treatments to test whether what people say they would do with what they actually do when given an opportunity. By comparing outcomes in these two contexts, the researcher can infer the presence of hypothetical bias, its causes and relative effectiveness of different mitigation techniques. Moreover, other researchers

can replicate, and perhaps extend the experiment to test its robustness. Valuation experiments have a number of common elements. Subjects are often university students, but a substantial and growing number of valuation experiments are conducted in the field with non-students. Subjects are asked about their value for a specified good. Most experiments focus on an individual's willingness-to-pay, although there are a few studies that also investigate willingness-to-accept (e.g., Bishop and Heberlein 1979 ; Coursey *et al.* 1987; Brookshire and Coursey 1987; Smith and Mansfield 1998; List and Shogren 2002). An important distinction among experiments is the type of good valued—some studies use private goods, others public goods.

Although the goal of nonmarket valuation is to value public goods, a great many experiments rely on private goods, including perhaps the majority of experiments we discuss in this chapter. With both private and public goods, the subject is required to value an exogenously provided commodity. Unlike many public goods, respondents are often familiar with private goods and their substitutes, and may have considered engaging in a market transaction for the private good in question at some point. Even if the good is unfamiliar, subjects may be more comfortable with valuing private goods, whereas the subject may have never considered placing an economic value on a public good like atmospheric quality. If subjects are more comfortable valuing goods they commonly purchase, then they may be less prone to error (List and Gallet 2001). If CV cannot accurately estimate economic value in these relatively familiar settings, it is probably unlikely to do so with public goods. Private goods also avoid any biases due to free-riding.

The prototypical economics experiment uses values induced by the experimenter (Smith 1976). Most valuation experiments, on the other hand, do not use induced values.⁵ Instead, the researcher tries to measure a respondent's subjective, homegrown values (Harrison *et al.* 2004 discuss some methodological issues that need to be considered when eliciting subjective values). As with an actual field CV study, these homegrown values cannot be known with certainty. However, by carefully manipulating the conditions under which values are elicited, the experimenter can test whether changes in explanatory variables influence responses.

2. Hypothetical Bias in Contingent Valuation

The larger debate over whether hypothetical responses reasonably approximate real responses gained great salience from claims for damages to public resources stemming from the

1989 *Exxon Valdez* oil spill. This claim, based on a stated preference study of passive use values, was vastly larger than other damage claims under CERCLA.⁶ The defendant in this case, Exxon Corporation, had very deep pockets and could fund its defense with munificence. The size of the initial damage claim, involving real resources, and the ability of Exxon Corporation to fund its defense led to an intensive study of the method. Both the critique of contingent valuation⁷ and the muted defense of the method by the NOAA Blue Ribbon Panel were a product of the Exxon Valdez oil spill. While all of major issues surrounding the use of stated preferences were debated, the central issue that arose then and remains today is the correspondence between real and hypothetical measures of economic value.

Because the CV practitioner cannot observe actual behavior, there is no way of knowing whether survey responses are consistent with what the respondent would do if actually given the opportunity—and there is plenty of evidence dating back to at least the 1930s that stated intentions can differ significantly from observed actions (LaPiere 1934; Schuman and Johnson 1976; Ajzen *et al.* 2004).⁸ The difference between stated and observed actions is known as hypothetical bias. The presence of hypothetical bias has been well-documented in both laboratory and field studies. Meta-analyses of the related experimental literature by List and Gallet 2001 and Murphy *et al.* 2005a report that mean hypothetical values are about two to three times greater than actual values (this comes from a highly skewed distribution with a median closer to 1.5). However, unlike the WTA/WTP disparity, hypothetical bias is not a behavioral anomaly and does not necessarily indicate a deviation from neoclassical demand theory. CV surveys are hypothetical in both the payment for and the provision of the good in question and economic theory provides no guidance about choices that lack salient economic consequences. Nevertheless, since the values in many cases may be substantial, and estimates of these values can influence whether a policy decision is in favor of environmental protection or development, they cannot be ignored. Of course, this raises the question of how to develop a mechanism that elicits unbiased responses or accurately calibrates biased value estimates.⁹ This requires a better understanding of why stated intentions differ from actual behavior. In this chapter, we discuss the experimental literature on hypothetical bias and conclude that future experimental research should focus more on developing a conceptual framework for understanding the underlying causes of hypothetical bias. Until then, attempts calibrate CV responses will continue to face questions about the robustness and generalizability of the approach.

2.1 The experimental approach to testing hypothetical versus real valuation

Researchers have been concerned about degree to which hypothetical responses approximate real responses since economists first attempted contingent valuation. The first tests of hypothetical versus real were the field experiments by Bohm and by Bishop and Heberlein. The appeal of the experimental approach was the ability to control more elements of the valuation process than the field experiments allowed. Hence much of our evidence about hypothetical versus real comes from lab experiments. The core of most hypothetical bias experiments is a pair of treatments that differ in only one dimension: whether payment for the good in question is consequential. The “hypothetical” treatments follow the same procedures as a standard field CV survey and people are asked what they *would* do if they were given an opportunity. In the “actual” or “real” payment treatments, respondents are asked to make actual payments (or accept payment in the case of WTA) and the good is provided if sufficient funds are raised. Additional treatments may test the effectiveness of various approaches to align value estimates in the hypothetical and real treatments, e.g. cheap talk or uncertainty adjustments discussed later in this paper.¹⁰

Because the experimenter is usually eliciting unknown homegrown values, it is important to emphasize the inferences that can be drawn from the results. If hypothetical values exceed actual values, as is typically the case, then the data clearly support the argument that these values differ. However, without knowing the true economic value of the good, we make the reasonable *assumption* that the responses in real settings represent the true economic value, and therefore the hypothetical values must be overstated. Based solely on what can be inferred from the data, it is entirely possible that the converse is true, that the hypothetical values are accurate and the real values misstated (Harrison 2002).

2.2 Documenting the presence of hypothetical bias

The early experimental CV research focused on simply testing whether hypothetical bias existed. The primary focus of Bohm’s (1972) seminal work was not the hypothetical bias problem per se, but rather the extent to which strategic behavior, manifested as free-riding, existed under different real and hypothetical payment schemes for viewing a closed-circuit television program. His results suggest the presence of a modest amount of hypothetical bias.

The implications of Bohm's result for contingent valuation were first examined by Bishop and Heberlein 1979. This field experiment asked hunters who had won a free goose hunting permit through a lottery about their WTA to sell the permit. Individuals in the real payment treatment received actual cash offers to buy their permits for varying amounts between \$1 and \$200. Offers were presented as a dichotomous choice decision about whether to accept an enclosed check. Mean WTA in the real payment treatment was \$63 per permit. In the hypothetical treatment, mean WTA was \$101, suggesting that responses to a hypothetical survey overstated WTA by 60%.¹¹

The presence of hypothetical bias reported in both Bohm 1972, and the series of hunting permit experiments by Bishop and Heberlein (Bishop and Heberlein 1979, Heberlein and Bishop 1986) were controversial. Mitchell and Carson (1986, 1989) re-examined these studies and found no hypothetical bias. They argued that Bohm's result was unduly affected by an outlier, and that the results of Bishop and Heberlein's goose hunting experiment were sensitive to the way in which unreturned surveys were interpreted. Hanemann 1984 also disputes the conclusion about hypothetical bias in Bishop and Heberlein 1979 by highlighting the sensitivity of the results to modeling assumptions.¹²

Dickie *et al.* 1987 went door-to-door offering to sell pints of strawberries and failed to observe any hypothetical bias, and Sinden 1988 also does not observe hypothetical bias using a within-subject comparison. Interestingly, after a critical examination of both Bohm, and Bishop and Heberlein, Mitchell and Carson 1989 uncritically point to the Dickie *et al.* study as evidence that CV can provide an accurate prediction of real market behavior. However, Harrison and Rutström forthcoming, suggest that a more detailed examination of their results yields mixed conclusions, and note that on average hypothetical responses exceeded actual payments by 58%. Moreover, Harrison 2002 notes that although the hypothesis of no bias was rejected at the 1% level, it is reject at the 1.2% level. Harrison also cites an unpublished manuscript by Hausman and Leonard which found that the hypothesis tests were calculated incorrectly, and that the hypothesis is rejected even at the 1% level.

Subsequent studies in the late 1980's had mixed results about the presence of hypothetical bias. Brookshire and Coursey 1987 found substantial bias in the WTA for changing the tree densities in a Colorado park (hypothetical responses were about 25 times greater than actual payments). The hypothetical bias they observed for WTP was more consistent with that

reported in other studies—about double. Coursey *et al.* 1987 also find a large disparity between WTP and WTA to taste a bitter substance (sucrose octa-acetate). The hypothetical bias in WTA is about double, but they find no discernible difference between hypothetical and actual WTP estimates. The elicitation procedures differ between real and hypothetical treatments, and this could confound interpreting results; they asked open-ended questions to elicit hypothetical values, but then held an n^{th} price Vickrey auction in the real treatments. Kealy *et al.* 1988 observe a bias when subjects were asked about their WTP for chocolate bars.

2.3 Elicitation procedures

Hoehn and Randall 1987 demonstrate theoretically that a dichotomous choice elicitation mechanism is incentive compatible, whereas an open-ended format is can lead to under-revelation of WTP (assuming, of course, that actual payments will be made). In the 1990s, much of the literature focused on testing whether different response formats were incentive compatible, and the extent to which this could influence conclusions about hypothetical bias. Neill *et al.* 1994 also recognized the potential for confounding effects if the elicitation procedures between the real and hypothetical payment treatments differ (as was the case in many early studies). Subjects were asked about their WTP for a print of a 16th century world map. The study consisted of three open-ended response formats: a real payment Vickrey auction, a hypothetical Vickrey auction and a hypothetical CVM in which subjects were asked an unstructured open-ended question about their maximum WTP. They report two key results. First, they find no significant difference in WTP between the two hypothetical treatments, which they suggest implies that hypothetical bias is more likely a function of whether payments are real rather than the elicitation mechanism. Second, values elicited in the hypothetical Vickrey auction are at least three times larger than when payments are real even though in both treatments subjects were informed of the demand revealing properties of the mechanism.

Cummings *et al.* 1995a extended the work of Neill *et al.* 1994 by asking whether a dichotomous choice response format is incentive compatible when eliciting WTP for three different common private goods (calculator, juicer, chocolate). In all cases, substantial hypothetical bias was observed. Brown *et al.* 1996 essentially combined the open-ended format in Neill *et al.* 1994 with the dichotomous choice format in Cummings *et al.* 1995a. Using a 2-by-2 design that crosses the response format (open-ended or dichotomous choice) with payment

type (hypothetical or real) to elicit WTP for road removal in the Grand Canyon, they observe hypothetical bias with both response formats. Moreover, mean WTP differs between response formats when payments are hypothetical. However, with real payments the differences between response formats are not significant. They suggest that the hypothetical nature of CV amplifies the difference in response formats.

Cummings *et al.* 1997 use a similar approach to test for hypothetical bias in a majority vote referendum. Subjects voted on whether everyone would contribute to printing a citizen's guide to groundwater for low-income families in New Mexico. The frequency of yes votes was about 67% higher in the hypothetical treatment than with real payments. Taken together, this series of studies suggests that hypothetical bias is likely to persist across all three response formats (open-ended, dichotomous choice and referendum).

Cummings *et al.* 1997 uncovered the issue of conditional versus marginal distributions of valuations. This is an unusual result because of the random assignment in experimental studies. Cummings *et al.* tested their hypothetical versus real result by estimating a probit for whether the subject is willing to pay, with a dummy variable for the real treatment. The exogenous variables included other aspects of the calibration procedure as well as seven socioeconomic characteristics of the subjects. They found that the effect for the 'real' response was significantly less than zero, supporting the hypothesis that hypothetical response exceeds the real. In a response to this paper, Haab Huang and Whitehead re-estimate the probit, arguing that the greater uncertainty about the meaning of a hypothetical question would warrant allowing more dispersion. When they permit the variance of the error terms to differ between the real and the hypothetical, they cannot reject the hypothesis that the real and the hypothetical are the same. One reason for the failure of randomization in this case may be that all subjects were given the same price of \$10—i.e. there was no randomization with respect to price.

2.4 Calibration Techniques

One of the outcomes of the NOAA blue ribbon panel was the effort to calibrate real and hypothetical responses. In particular, the NOAA panel noted the “unfortunate” lack of data that could be used to calibrate CV responses (p. 52). In the absence of a means to calibrate responses, the panel recommended that CV responses be divided in half. This helped spawn a

shift in the focus of the experimental literature from testing whether hypothetical bias exists to developing calibration techniques that could mitigate this bias.

These calibration techniques come in two forms: (1) *ex ante* attempts to elicit unbiased responses through survey design (also referred to as instrument calibration, see Harrison (2006)), and (2) *ex post* methods that attempt to calibrate biased responses (statistical calibration).

2.4.1 *Ex ante* approaches or instrument calibration

The NOAA panel hypothesized that in a hypothetical survey respondents might not carefully consider their budget constraints, but if they were given a reminder about these constraints and opportunity costs, then people would revise their responses downward. The panel reasoned that because CV surveys lack salient economic consequences, people might not carefully consider their disposable income, whereas an actual solicitation for contributions would force this to play a role in the decision. As a result, the panel recommended that CV surveys include convincing reminders of economic constraints and the availability of substitutes. In effect, the panel was advocating an instrument calibration approach in which the survey is designed to elicit unbiased responses *ex ante*. Loomis *et al.* 1994 tested this hypothesis with a simple design that used two similar survey instruments to elicit hypothetical willingness to pay for a reduction in fire hazards to old-growth forests: one with a reminder about budget constraints and substitutes, the other without a reminder. Contrary to the NOAA panel's expectations, they found no difference in mean willingness to pay between the two treatments. Since their design did not include a treatment with real economic commitments, the absence of any changes in mean WTP with the budget reminder has two possible interpretations: (1) the reminder was ineffective at eliminating hypothetical bias; or (2) the hypothetical responses were already consistent with what people would actually do, and there was no bias that needed to be corrected in the first place. By including a real payment treatment, Loomis *et al.* 1996 did not have this interpretation problem when they tested a hypothetical CV survey which combined a budget reminder with an explicit request to "...answer as if it were real—as if you were participating in a real sealed-bid auction and would really have to pay your dollar amount if you were the highest bidder" (p. 453). This combined reminder/request did reduce WTP estimates for an art print, but not enough to eliminate hypothetical bias altogether.

A budget reminder is only one type of corrective entreaty which attempts to tackle the hypothetical bias problem indirectly by making an assumption about its underlying cause and asking respondents to account for this in their decisions. Cummings *et al.* 1995b, describe the results of a practical and intuitively appealing technique for eliciting unbiased responses *ex ante*: simply make people aware of the hypothetical bias problem and to account for it when making their decisions—the cheap talk idea. They extended the design in Cummings *et al.* 1997 by comparing a real payment treatment with three hypothetical treatments: a baseline plus two versions of a cheap talk script, which they referred to as “light” and “heavy.” The former was an abridged version of the latter. The scripts were read to student subjects immediately before a referendum that would require everyone to pay \$10 to help finance the publication of a citizen’s guide to groundwater quality. The scripts described the hypothetical bias problem, summarized the results of Cummings *et al.* 1997, and encouraged respondents to think carefully before voting. The main difference between the two scripts was that the heavy script was longer, provided more details about the Cummings *et al.* 1997, study, and offered some conjectures about why the hypothetical bias problem exists.

In principle, because the cheap talk script does not directly affect the payoff structure of the game, it should have had no effect on behavior—but it clearly did. However, their results about its effectiveness in eliminating hypothetical bias were mixed. The light (or short) script actually *worsened* hypothetical bias—the probability of a yes vote in the referendum increased by 21%. The heavy (long) script, on the other hand, successfully eliminated hypothetical bias by bringing hypothetical responses in line with those from the real payment treatment. Cummings and Taylor 1999 subsequently tested the robustness of the positive results from the heavy cheap talk script across three commodities: the same citizen’s guide to groundwater, plus two referenda about donations to Nature Conservancy for land preservation, one in Georgia and the other in Costa Rica.¹³ Consistent with the earlier results of Cummings *et al.* 1995b, they found that the heavy cheap talk script was effective in eliminating hypothetical bias for all three goods.

The basic premise behind cheap talk is that simply making respondents aware of the hypothetical bias problem is sufficient to make the CV instrument demand revealing. While the rationale may seem intuitive, the absence of a comprehensive theoretical framework to explain why hypothetical survey responses differ from observed behavior opens up many questions about why cheap talk is effective and the conditions under which it might succeed or fail.

Shogren 2005, suggests that “Perhaps there is some deep cognitive reason or maybe it is just the Hawthorne effect at work—subjects want to meet the expectations of the experimenter” (p. 1010). It is also possible that the long cheap talk script is confounded with a change in the respondent’s perception of the good because the script implies that the good may be more valuable than the individual initially thought (Harrison 2006). An advantage of the experimental method is that these hypotheses can be tested. Experiments facilitate not only the development of new techniques, but also the testing of their effectiveness across a wide array of variables. Hence, it should not be a surprise that the promising results in Cummings and Taylor 1999, has led to a plethora of studies examining its robustness. To date the results are inconclusive, but some common patterns do seem to be emerging.

In a field study that elicits WTP for a baseball card using both card dealers and non-dealers, List 2001, finds that Cummings and Taylor’s long cheap talk script does eliminate hypothetical bias for non-dealers, but has no effect on the inflated hypothetical bids of experienced traders. He hypothesizes that those who are familiar with the good have well-structured preferences and are therefore less likely to rely on external signals, such as a cheap talk script, when determining their bids. Both Aadland and Caplan 2003, and Lusk 2003, report similar findings. Landry and List 2007 also conducted a field experiment at a sports card show but used only inexperienced subjects. Consistent with the previous results about trading experience and familiarity with the good, the long script successfully aligned hypothetical and real WTP. And, in a field study that elicited WTP for a pharmacist-provided diabetes management program, Blumenschein *et al.* 2008, report that the long script had no effect on WTP. The public is generally unfamiliar with such programs, but the study participants were all diabetics, who were therefore more likely to be experienced and well-informed. Although the designs of last two studies do not facilitate a direct comparison of the effectiveness of long cheap talk script between experienced and inexperienced groups, their results are consistent with studies that do.

There is some evidence that the long script may be more effective at higher payment amounts. Cummings and Taylor only used a single payment amount (\$10), whereas Brown *et al.* 2003, held referenda for amounts between \$1 and \$8. They found that cheap talk eliminated hypothetical bias associated with \$5 and \$8 payments, but had little effect on lower amounts. Subsequent work by Murphy *et al.* 2005b and Whitehead and Cherry 2007, use wider ranges of

payment amounts and also find that cheap talk is more effective with higher prices. In a study that uses induced values, Aadland *et al.* 2007 find cheap talk to be weak but effective for high amounts, but has no effect at lower announced prices.

Cummings and Taylor acknowledge that their long cheap talk script might be impractical for telephone surveys and recommend further study to determine the minimum effective script length, particularly in light of their earlier results with the short script which exacerbates the bias (Cummings *et al.* 1995b). Aadland and Caplan 2006, get a similar negative result using a neutral short script that intentionally avoids mentioning the direction of the bias by telling respondents that studies have shown that people tend to misstate (as opposed to overstate) their values. On the other hand, Poe *et al.* 2002, find that a script that was even shorter than that of Cummings *et al.* 1995b, has no effect on hypothetical bias. The short script does eliminate hypothetical bias for Aadland and Caplan 2003, but only for those households with strong environmental preferences. The short script also reduces hypothetical WTP in both Bulte *et al.* 2005, and Whitehead and Cherry 2007, but in the absence of a real payment treatment, it is impossible to determine whether the short script completely eliminates hypothetical bias (or possibly even over-corrects for it).

The primary goal of corrective entreaties like cheap talk and budget reminders is to get people to respond to hypothetical surveys *as if* their decisions had salient economic consequences. An alternative *ex ante* approach is to convince respondents that there is at least a chance that their responses would actually have real consequences, such as by providing survey results to policy makers. Carson and Groves 2007, make the case that respondents must believe that the survey results could influence decision makers and ultimately affect outcomes. If, in addition, the individual has preferences over the set of outcomes, then they suggest that the survey instrument is consequential and capable of inducing truthful responses. Dillman 1978 makes a similar point when he emphasizes that respondents should be informed about the social usefulness of the study. This, of course, raises the question of how realistic or credible the survey instrument must be to induce truthful responses. Cummings and Taylor 1998, provide some insights into this question using an experimental design that varies the probability that a survey referendum will result in a binding economic commitment. In addition to the standard hypothetical and real treatments in which the participant knows with certainty whether the referendum results will be binding (0% and 100%, respectively), they also included three

probabilistic treatments: 25%, 50% and 75%.¹⁴ Their goal was to find the minimum probability such that hypothetical bias was no longer present. As the probability of the referendum having consequences increased, the percentage of subjects voting in favor consistently decreased and approached that of the real payment treatment. However, only the treatment with a 75% probability was able to elicit responses that were statistically indistinguishable from the real payment (100%) treatment.

Whereas Cummings and Taylor found that a 50% probability was insufficient to induce truth telling, Landry and List 2007 report that a coin toss to determine whether the referendum results would be consequential did successfully align hypothetical and real responses. Although the results of the two studies are not directly comparable as they differ across important dimensions (e.g., students vs. non-students, open vs. closed referendum, public vs. private good), together they do provide some preliminary evidence that increasing the respondent's subjective assessment of the probability of consequences might lead to improved value estimates. The important question, however, is not whether the correct probability for accurate calibration is 50 or 75 percent, but rather how consequentialism could be implemented in the field where probabilities are subjective and not known by the experimenter. In a field study by Bulte *et al.* 2005, respondents were simply told that "the results of this study will be made available to policy makers, and could serve as a guide for future decisions with respect to taxation for this purpose" (p. 334). This statement was sufficient to reduce hypothetical WTP and equate it WTP estimates from a hypothetical cheap treatment. However because the good valued is a public good—government actions to protect seals in the Netherlands—the study lacks the real payment treatment necessary to evaluate its overall effectiveness.

The *ex ante* devices of cheap talk, budget reminders and similar measures are aimed at making the hypothetical responses more closely aligned with real responses. The literature also provides a parsimonious approach to determining how 'real' the subject believes her answer to be, by asking how certain the subject is of her response. CV studies began asking subjects about the likelihood that their answers were accurate well before the responses were used in the context of hypothetical versus real. Champ, Bishop, Brown and McCollum (CBBM) first used questions such as these to calibrate hypothetical versus real responses. In their mail study of the willingness to pay for removing old roads on the North Rim of the Grand Canyon, CBBM ask

for real and hypothetical donations. They then show that those subjects who are certain have hypothetical responses that are similar to real responses.

This result was replicated in an experimental setting by Blumenschein et al. (1998). In a lab setting, they offered subjects the opportunity to purchase a special pair of sunglasses. After the hypothetical question, they asked the subjects who answered yes whether they were ‘probably sure’ or ‘definitely sure’, and counted as yes responses only from the ‘definitely sure’ subjects. Unconditional proportions were almost identical for the ‘definitely sure’ yes hypothetical responses and real responses at a low price while the proportion of all hypothetical responses is significantly larger than the real response. At a high price, there is no difference between hypothetical and real.

Blumenschein et al. (2008) continue this line of investigation with a field study valuing a new and innovative diabetes management program with subjects who were diabetics. In this case, they compared a cheap talk script with the same two categories of expressed uncertainty: probably sure and definitely sure. The program is offered to 90 subjects as a real purchase; 187 subjects are offered the program hypothetically, of whom 86 receive a cheap talk script. The elicitation method is a dichotomous choice with random price. All of the subjects in the hypothetical group are asked the question about how sure they are. Since this comes at the end of the session, it cannot contaminate the hypothetical responses. They find that proportion of the hypothetical group (excluding the cheap talk group) willing to buy the program including all of the yeses is significantly greater than in the real group. When the hypothetical group includes only the ‘definitely sure’ responses, the proportions are not statistically different from the real response. When the hypothetical with cheap talk group is compared with the real group, the responses are higher than the real group at high prices and when the proportions are taken across all prices. This result on hypothetical responses seems especially convincing because the subjects are diabetics who have strong incentives to be knowledgeable about the diabetes control program and would seem less susceptible to hypothetical bias than in more ordinary circumstances where subjects value sunglasses or sports cards or mugs. The weight of the evidence in Blumenschein et al. (2008) supports the use of a certainty measure over cheap talk to control for hypothetical bias.

2.4.2 *Ex post* or statistical calibration

The idea of ‘adjusting’ a hypothetical response function or value arose during the development of the NOAA Blue Panel report. This idea, first explored by Blackburn, Harrison and Rutström, (BHR) envisions predicting the bias from a hypothetical willingness to pay function. As BHR explain, the hypothetical responses may give substantial overestimates of real WTP but ‘The hypothetical responses can still be informative as to real responses if the bias between the two is systematic and predictable’ (p. 1084). BHR find that discrete choice offers to purchase a good hypothetically substantially exceed those to pay real money for the same good. They estimate a multinomial logit model as a function of individual characteristics, where the alternatives are yes to hypothetical-yes to real; yes to hypothetical-no to real; no to hypothetical-no to real. While the fit is not strong, a model like this under the proper circumstances could be used to predict the probability that an individual with known characteristics who responded yes to a hypothetical CV would respond yes to a real question.

List and Shogren (1998) approach the statistical calibration problem more directly. In a field experiment, they run three types of experiments in which they obtains values with a Vickrey second price auction. Each subject participated in a hypothetical auction followed by a real auction. They are then able to estimate models in which the real valuation is estimated as a function of the hypothetical valuation. Presumably these results could be used for other valuation tasks. While the List-Shogren results show that hypothetical valuation significantly exceeds real values, they also find that the equations predicting real valuation as a function of hypothetical valuation are different for different goods and different subjects. They recognize the ‘problem of calibration of non-deliverables’—can we use calibration functions estimated in the lab or in field experiments for correcting hypothetical bias for goods for which stated preferences are necessary?

In the end, there are two problems with calibration functions. As BHR indicate, it suffers the same problems as benefit transfer—when we estimate valuation for a good or resource in one setting and apply it to another, it is unlikely that the two situations are the same, and that real values would be the same. The problem is heightened for hypothetical valuation, however because one might argue that values ought to be similar for similar situated resources, but there is no strong reason for hypothetical bias to be systematic in the same way. Without a great more evidence on how calibration actually works, we have little basis for arguing that biased hypothetical responses can actually be corrected statistically.

3. Willingness to Accept and Willingness to Pay

The divergence between willingness-to-accept and willingness-to-pay was the first experimental anomaly in the field of environmental economics, and among the first to show up in general economics. Concerns that the observed disparities posed a challenge to the neoclassical model surfaced immediately. While challenges to the neoclassical model have multiplied over the years – it is almost impossible to count them – the relationship between willingness-to-accept and willingness-to-pay remains instructive and pertinent.

3.1 Model

The ideas of willingness-to-accept and willingness-to-pay are rooted in neoclassical welfare economics. Willingness-to-accept (WTA) is the amount of money that would make an individual indifferent between a particular bundle of goods and a different, lesser bundle of goods plus the extra money. Willingness-to-pay (WTP) is the amount of money that would make an individual indifferent between the lesser bundle of goods and the greater bundle of goods but absent that amount of money. We refer to experiments that elicit these values as valuation experiments. In some contexts, WTP refers to aggregate or mean willingness-to-pay over a group of individuals; in other contexts, WTP refers to an individual-specific quantity. The context should make clear which definition is being referred to.

Typically, the difference between the initial bundle of goods and the alternative bundle is the quantity of just one of the goods. That good may be a private good, such as a mug or flashlight, or a public-type good such as air quality, access to a fishing site, or more acres of forest. The good in question should be “rationed” or exogenous which means that the individual cannot purchase it or otherwise directly affect how much of it he consumes or experiences.¹⁵ This assumption is not always clear or valid. The literature frequently refers to the rationed good as a public good.

For public goods, the assumption that the good’s quantity is exogenous to the individual is reasonable and natural. Sometimes this property is stretched or violated, however, as when the good involves both a “true” public good (air quality) and a quasi-public good, such as trips to visit a forest, which is a private good with a public good as an attribute. In the majority of experimental settings the good is actually a private good. When the experiment involves private

goods that are available outside the experiment (mugs or flashlights), individuals are presumed to focus attention narrowly around the experiment, during which time these goods are indeed beyond control of the subject. In making this presumption, WTA/WTP experiments are no different from other economic experiments, such as those eliciting risk attitudes.

The neoclassical model starts with the primary utility function defined over a vector of goods x purchased at prices p , a rationed good q , and income, y . Indirect utility is then given by:

$$(1) \quad V(q, y) = \max u(x, q) \text{ subject to } px = y$$

The vector p is usually suppressed as an argument of the indirect utility function in this context. See Bockstael and McConnell for an analysis of the properties of utility function in the presence of an exogenous good.

WTA and WTP are defined implicitly as the solutions to the following two equations:

$$(2a) \quad V(y, q_1) = V(y + WTA, q_0)$$

$$(2b) \quad V(y, q_0) = V(y - WTP, q_1)$$

where $q_1 - q_0 = \Delta q > 0$. Δq is what is being valued in each of the set-ups. It is conceptually possible to define willingness-to-pay or willingness-to-accept in terms of another rationed good (“How many additional candy bars would make you indifferent to the loss of a movie ticket?”). The special insights of these experiments are discussed below. The definitions in (2) also assume that q can be varied independently of both p and y . There is no uncertainty.

Analysts sometimes mistakenly substitute:

$$(2b') \quad V(y, q_1) = V(y - WTP', q_1 + \Delta)$$

for (2b) and use the results to compare WTA with WTP' . This substitution is probably not important empirically, however, since WTP' is usually almost exactly equal to WTP in experiments.

3.2 Evidence

The research that first brought WTA and WTP to the profession's attention was a study of waterfowl hunting by Hammack and Brown. Their study, one of the earliest contingent valuation studies, provided a within-subject comparison of the two values. When the authors omitted protest bids (12.4 percent for WTA and 1.4 percent for WTP), their estimate of mean WTA (to give up duck hunting for one season) was \$1044 and mean WTP (for the right to hunt for one season) \$247. This is a ratio of about 4.3, which was much larger than the authors expected. Thus it set the stage for the study of WTA vs. WTP. In retrospect, the ratio uncovered by Hammack and Brown was actually *smaller* than should be expected; Horowitz and McConnell (2002) found that the average WTA/WTP ratio for hunting licenses (including Hammack and Brown's finding) was roughly 10. Attention originally focused on the special role played by environmental goods, although researchers soon after learned that the WTA-WTP disparity existed across a wide range of goods.¹⁶

Over the years, systematic evidence of unusually high WTA/WTP ratios has built up. The evidence comes from a variety of sources: stated preference studies conducted expressly for addressing policy design or resource allocation (Rowe, d'Arge and Brookshire), lab experiments meant to understand or explain the gap (Brookshire, Coursey, and Schultze; Knetsch 1989), and field experiments (Bishop and Heberlein; List 2004). Non-laboratory revealed preference studies have been rarer. Zeckhauser and Samuelson provide extensive evidence of preference for the status quo, which is related to the WTA-WTP phenomenon; a good deal of their evidence can be considered revealed preference but does not directly provide estimates of WTA-WTP differences. Chattopadhyay (2002) is the only study that we know of that finds more than rounding error differences in WTA and WTP by estimating a neoclassical preference function rather than separate experiments for WTA and WTP as in (2a) and (2b). Typically estimates of WTA and WTP induced from revealed preference methods find no significant WTA-WTP gap.

Experiments typically assign participants randomly into two groups, one of which receives some good and one of which does not. Some studies keep the participants in the same room and randomly assign the good to some of the subjects, who are then asked their WTA; subjects who did not receive the good are asked their WTP. More often, studies keep the participants in the two treatments in separate rooms so that they do not see the other "state of the world." Given random assignment, the mean of WTA from one group should equal mean WTP from the other group in the absence of an anomaly or a substantial income effect. The ratio of

mean WTA to mean WTP in the two treatments is what is most commonly reported and discussed and is simply referred to as WTA/WTP. Some studies compare group medians. A few studies, including Hammack and Brown, asked all participants for both *WTA* and their *WTP*. While with-subject treatment is not regarded as good experimental practice, it is not obvious given the nature of the questions that there would be unobserved correlation in these responses.

None of these experimental treatments appears to affect the basic finding of WTAs substantially higher than WTPs. Two studies that we know of reported both within-subject and among-subject mean ratios and found that the mean WTA/WTP ratio was actually higher than the ratio of mean WTA to mean WTP. We do not know of any study that empirically estimates the effect of protocol, such as whether it matters if individuals see or know about participants in the other treatment.

Indeed, one of the findings of the WTA/WTP literature is that the ratio is roughly immune to survey design. It just does not seem to matter very much how the experiment is conducted, within bounds. Some exceptions are discussed below.

Horowitz and McConnell (2002) summarized the evidence on the *WTA-WTP* ratio as of 2002. Their main analysis examined 45 studies with 201 observations on *WTA/WTP* ratios. They found a mean ratio of 7.2, with individual ratios ranging from 0.74 to 113, albeit with many higher ratio findings excluded from the analysis. The goods being valued had a mean WTP of \$175 (using 1983 dollars) and a median WTP of \$3.73. Of greater interest is the wide range of goods that have been studied, both in the Horowitz-McConnell review and subsequent research: chocolates, pens, mugs, movie tickets, baseball cards, hunting licenses, potted plants, visibility, nasty-tasting liquids, pathogen-contaminated sandwiches, acres of preserved habitat.

Horowitz and McConnell showed that farther the good is from being an ordinary market good, the higher the ratio. This finding was robust to the fineness of the classification scheme (how ordinary is “ordinary”?) and survey design. This is an intriguing and informative finding but at the same time it has not provided the breakthrough in behavioral economics that one might expect from such clear results: The observed pattern is consistent with a great many alternative theories of economic behavior.

By and large, researchers have agreed that the ratios are larger than one might expect intuitively. These ratios are the empirical evidence that require explanation. The explanations of these ratios are of four sorts: (i) more rigorous application of the standard neoclassical model, (ii)

expanded models that are still within the neoclassical paradigm, (iii) models that are not neoclassical (that is, behavioral models), and (iv) experimental artifacts that do not appear in better designed experiments or in real world behavior.

3.3 Tests of the Neoclassical Model based on WTA and WTP

The debate about whether the *WTA-WTP* ratios are consistent with neoclassical demand began with Willig's (1976) paper on bounds for equivalent variation and compensating variation for price changes. Willig showed that one could calculate bounds for equivalent and compensating variation for price changes using information commonly available for marketed commodities. Willig summarizes his results in a table that shows that *WTA* and *WTP* typically differ from consumer surplus by a small percent, and hence from each other by a small percent. Only when income elasticities and the ratio of consumer surplus to income are relatively high (not typically observed) will there be large gaps between *WTA* and *WTP*.

The Willig results are suggestive but do not apply to the typical valuation case in which the quantity of the (rationed, free) good is varied rather than price. This distinction was first noted by Randall and Stoll who developed bounds similar in spirit to the Willig bounds but with a model of exogenous quantities. The Randall and Stoll bounds were similar to the Willig results, showing small differences between *WTA* and *WTP* for reasonable preference parameters. They also showed that when the Marshallian surplus is very large and the income elasticity of marginal willingness to pay also large, the *WTA/WTP* ratio can become quite large and still be consistent with neoclassical preferences. In other words, like subsequent researchers, they recognized that there were possible specifications of neoclassical preferences that would be consistent with the large *WTA/WTP* ratios.

Under the model in (1) and (2), *WTA* and *WTP* should be close when Δq is small. One way to see this is to use a Taylor series on the expressions in (2) to derive the approximations shown in (3). Care must be taken with the approximations because both approximations must go in the same direction to be comparable. Since *WTA* and *WTP* experiments measure required changes in income, it is necessary to construct series that start with the same q 's and move in the income dimension. This approach yields:

$$(3) \quad WTA \approx \frac{V_q(y, q_1)}{V_y(y, q_1)} \Delta q \quad \text{and} \quad WTP \approx \frac{V_q(y - WTP, q_1)}{V_y(y - WTP, q_1)} \Delta q$$

The difference is then:

$$(4) \quad WTA - WTP \approx \frac{V_q(y, q_1)}{V_y(y, q_1)} - \frac{V_q(y - WTP, q_1)}{V_y(y - WTP, q_1)} = -WTP d\left(\frac{V_q}{V_y}\right)/dy$$

The latter equality arises from a Taylor series expansion of V_q/V_y with the functions evaluated at some point in the interval between $\{y - WTP, q_1\}$ and $\{y, q_1\}$.

The derivative term is:

$$(5) \quad d\left(\frac{V_q}{V_y}\right)/dy = \left(\frac{V_{qy}}{V_y} - \frac{V_q}{V_y} \frac{V_{yy}}{V_y}\right)$$

Suppose the initial Taylor expansions are valid. Then WTA will be close to WTP whenever this expression is small. V_{yy} can be set equal to zero without loss of generality. Therefore WTA greater than WTP requires $V_{qy} < 0$. This condition means that having a higher q reduces the marginal utility of money. A lower marginal utility of money implies that the individual needs a higher amount of money to compensate for the loss of q . In other words, money is less valuable in the WTA case (the individual starts with q_1) than in the WTP case (the individual starts with q_0). If money is less valuable, the person needs more of it to compensate for a change in the good q . This expression and the role for V_{qy} shows why WTA/WTP is often referred to as arising from an income effect.

This framework then shows why WTA should be close to WTP: We do not expect changes in consumption of the experimental good – whether a movie ticket, a hunting license, or visibility at a national park – to substantially affect the marginal utility of money.

Hanemann and Amiran and Hagen (AH) implicitly argue that researchers may be unrealistically expecting V_{qy} to be small. Both articles use substitutability to motivate their results.

Hanemann begins with intuition based on two polar cases — the Leontief utility function in which there is no substitution between the public good and any private good, and a utility function in which the public good is a perfect substitute for at least one good. The Leontief example is instructive even though extreme. Consider a version of the model with just two goods, hamburgers (h) and buns (b). Suppose the individual receives utility only when

hamburgers and buns are consumed together in a one-to-one form, $U = \min \{h, b\}$. Hamburgers have price p_h . Suppose further that buns are free but rationed. The individual purchases hamburgers up to the amount b^* . Suppose $p_h b^* = Y$, the individual's income; the individual spends his entire budget on hamburgers. Utility is equal to b^* . In this situation, his willingness to pay for an additional bun is zero because he cannot afford any more hamburgers to go with it; he is already spending all of his money. His willingness to accept the loss of a bun is infinite, however, or undefined. A reduction in b^* cannot be made up for by greater purchase of hamburgers, even with additional income, because buns, not hamburgers determine his utility. No amount of additional money, allowing him to purchase as many hamburgers as he wants, will allow him to reach the utility he previously had with b^* buns. No amount of money can compensate him for a reduction in b^* .

This example is extreme but any utility function with high complementarity (i.e., lack of substitutability) between the rationed good and all other goods will yield similar outcomes: A small willingness to pay and a high willingness to accept.¹⁷

In the second case, Hanemann shows that WTA equals WTP when there is perfect substitution between the rationed good and at least one of the other goods. Hanemann then develops a general expression showing that the difference between WTA and WTP depends on ratio of the elasticity of substitution between the composite commodity of private goods and the public good to the more familiar income elasticity of demand. Hanemann also argues, sometimes explicitly and sometimes implicitly, that many of the goods for which high WTA/WTP ratios are observed are goods with few substitutes. The Hanemann result has great intuitive appeal because it provides analytical support for a large WTA-WTP gap for a unique public good, which would often be characterized by lack of available substitutes.

Shogren, Shin, Hayes and Kliebenstein (SSHK) conducted experiments to test the prediction that less substitutability will increase the WTA-WTP difference. Using Vickrey second price auctions, they derive WTA and WTP values for two types of goods: candy bars, which have many substitutes, and risk of infection from food-borne pathogens, which they claim would have few substitutes. They found, after a series of trials, that for candy bars WTA converged approximately to WTP. For pathogen-contaminated sandwiches, WTA and WTP did not converge. The WTA/WTP ratios for the pathogen-contaminated sandwiches ranged from approximately three to five, whereas the WTA/WTP ratios for candy bars were close to 1

(SSHK, Table 4). The authors claim that this finding shows that the Hanemann explanation is correct.

There are two related difficulties with Hanemann's argument. First, we cannot observe the elasticity of substitution for goods that are rationed. Therefore, the claim that candy bars have more substitutes than do health risks from sandwiches can only be asserted, not checked. Second, even if we accept this assertion there is no way to know whether the degree of substitutability is sufficient to yield the observed WTA/WTP ratios. Indeed, we might as well use the WTA/WTP ratio to measure substitutability as use substitutability to predict WTA/WTP. In other words, the Hanemann result is not sufficient to provide a parametric test of neoclassical theory. It does, in some circumstances, provide a comparative static test ("Are ratios higher for goods with fewer substitutes?") but even this is weak because: (i) the claim that a good is more or less substitutable than some other good cannot be established and (ii) other theories of behavior yield the same predicted relationship but are inconsistent with neoclassical optimization.

Amiran and Hagen (AH) take a different approach by focusing on a commonly assumed property for utility functions, unboundedness. They claim that utility functions are likely to be asymptotically bounded and show that for asymptotically bounded utility, willingness to accept can be substantially higher than willingness to pay. Unbounded utility implies that there exists an amount of any one good that can "compensate for the loss of nearly the entire quantity of all other goods" (AH, p. 458), an unlikely property. Asymptotically bounded utility does not exhibit this property. AH then demonstrate that with asymptotically bounded utility there is always a budget (income and price) that will lead to an infinite WTA for losses in the public good.¹⁸

As with Hanemann, AH's proposed utility function is consistent with low substitutability between the rationed good and other goods and with WTA substantially higher than WTP but it does not *calibrate* WTA/WTP ratios for specific goods. That is, it still does not provide a testable hypothesis for whether individual WTA/WTP ratios are consistent with the underlying utility model. While the Hanemann measure provides at least one property that can be approximated experimentally – the substitutability of the rationed good for other market goods – there are no obvious experimental procedures that follow from the AH model.

Although Hanemann and Amiran-Hagen are conceptually correct, they provide little assurance that observed WTA-WTP values are consistent with the model in (1). Since V_{qy} is

unobservable we cannot use the argument that it “could be large” as proof that it is actually large.

Tests of the validity of the neoclassical model must then take another approach. Horowitz and McConnell (2005), following Sugden, show that under model (2):

$$(6) \quad \frac{WTP}{WTA} \approx 1 - \frac{\partial WTP}{\partial y}$$

Since both $\frac{WTP}{WTA}$ and $\frac{\partial WTP}{\partial y}$ can be observed, this relationship can be tested.

Expression (6) shows that high WTA/WTP ratios imply that WTP must be very sensitive to income. For example, the Shogren et al. results for risk of food-borne infection gives ratios of WTA/WTP from 3 to 5, implying that $dWTP/dy$ ranges from 0.67 to 0.80. This in turn means that if an individual were given an extra dollar, her willingness to pay for the good would increase by \$0.67 to \$0.80. In other words, she would be willing to devote a large portion of an extra dollar of income to pay for the rationed good. Likewise, individuals with higher incomes should have substantially greater WTP. This implication is not surprising because a large income effect is precisely what is needed for large WTA/WTP ratios. The evidence is against it, however. While we do not have estimates of the marginal willingness to pay for avoiding food-borne risks, introspection suggests that it would fall far short of \$0.67—a very high proportion of income increases to devote to this risk.

Horowitz and McConnell (2005) examined the evidence for (6) and showed a number of very low estimates of $\frac{\partial WTP}{\partial y}$ for the goods and services that they surveyed. Although most estimates were zero or negative, their maximum $\frac{\partial WTP}{\partial y}$ was 0.0029 (Table 4), implying a WTA/WTP ratio of 1.003. Horowitz and McConnell (2005) undertake several other approaches for testing (6), including re-specifying the expression in terms of income elasticities.

The Sugden test and Horowitz and McConnell’s analysis are stronger than tests based on substitutability but still have weaknesses. First, one possibility is that utility is so sharply curved that the Taylor series approximation is not accurate. Second, many studies cannot reject the hypothesis that $\frac{\partial WTP}{\partial y}$ is zero and therefore do not report the estimated coefficient. Therefore,

researchers do not have as good a picture of $\frac{\partial WTP}{\partial y}$ as they might like. Of course, one reason that the estimated coefficient is often insignificant is that income is difficult to measure.¹⁹

These approaches to testing the neoclassical model are particularly odd because an exact test of the theory exists. This test is easiest to see in connection with reference dependence, a sort of umbrella alternative model under which individuals judge changes in q and income relative to a reference point. Under reference dependence, the indirect utility function can be written $\varphi(y, \Delta q; q)$. Utility depends on income and *changes* in q , with the value of changes in q possibly also depending on the endowment or reference point, q . (Note, in many cases, q is omitted and we write utility as a function solely of y and Δq .)

In the neoclassical model, the individual cares only about the final consumption of q and therefore q and Δq are *perfect substitutes*. Therefore we have the implication:

$$(7) \quad \frac{\partial WTP}{\partial q} = \frac{\partial WTP}{\partial \Delta q}$$

This is an exact test based on observable quantities. The experiments required to estimate $\frac{\partial WTP}{\partial q}$ and $\frac{\partial WTP}{\partial \Delta q}$ are straightforward, although they require substantially more observations than those needed to estimate, say, WTA and WTP. Note also that the test in (7) is based on WTP alone. It does not rely on combining WTA and WTP experiments.

Only one study that we know of directly tests (7). Horowitz, McConnell, and List estimated a utility function defined on trading ratios for baseball cards and found that a version of (7) indeed holds, but only if one also assumes reference dependence. In other words, equation (7) was found to be violated and neoclassical preferences rejected. See further discussion in Section 4.

3.4 Neoclassical Response

Economists have responded with more general models that are still within the neoclassical tradition yet are consistent with the empirical evidence. A separate set of papers have argued that the observed WTA-WTP disparities are an experimental artifact and therefore

do not provide reliable evidence against the neoclassical model. We examine each of these approaches.

3.4.1 Expanded models within the neoclassical framework

Zhao and Kling (2001) use a real options model to argue that WTP can be substantially smaller than WTA if an individual is uncertain about her value for the good and if there are “non-trivial transaction costs associated with reversing her purchasing or selling decision.” WTA is high because an individual might give up a good that she later learns is more valuable to her than she originally assumed, and can recover it only at a cost because of the transaction cost. WTP is low because an individual might purchase a good that she later learns is not so valuable; again, the transaction cost makes it costly for her to sell the good in this situation.

Zhao and Kling interpret several of the previous WTA/WTP studies in this light. Direct tests of the theory are in Kling, List, and Zhao (2003) and Corrigan, Kling and Zhao (2008). Kling, List, and Zhao (2003) conducted several sportscard auctions. In one experiment, subjects were asked about the ease with which they expected to be able to buy, sell, or trade the item after the auction. In a second experiment, which elicited only willingness-to-accept, subjects were given an option for purchasing the item (if they sold it in the auction and later decided they wanted it) or for selling it (if they did not sell it in the auction and wanted to try to sell it); these were compared to a control in which no such transaction-cost-reducing option was provided. In the first experiment they found that the higher the perceived difficulty of reversing the transaction the greater the WTA/WTP disparity. In the second experiment, the treatments to reduce subsequent transaction costs led to lower WTA than the control group. Both of these findings are consistent with the Zhao and Kling model. KLZ also found that WTA was substantially higher than WTP even for individuals who expected to trade the card ($\$10/\$7.67 = 1.30$). This ratio is roughly equal to the ratios found for this sort of goods by Horowitz and McConnell (2002) but still higher than the neoclassical model predicts (based on Horowitz and McConnell) absent transaction costs.

Kling and Zhao analyze the effect of delaying commitment in research on the value of water quality improvements. These findings show two weaknesses of the transactions cost explanation. First, WTA/WTP results can be generally consistent with the explanation but still not provide evidence for or against the neoclassical model, with option value and transaction

costs embedded. In other words, WTA/WTP ratios could be increasing in transaction costs or in the likelihood of future information about values and yet still be too high relative to observed income or substitution effects.

Second, this explanation depends on the likelihood that individuals in the future will have more information about the value of the item being studied. It is not clear that this kind of uncertainty is important for environmental issues. One of the examples cited by Corrigan, Kling, and Zhao (2006) concerns willingness to pay for a public park and the possibility that “at a future date residents may have a better estimate of the park’s value”. Why would they have a better idea of the park’s value in the future? In specific cases, perhaps, but in general, no. The claim sounds like hedging: Subjects have great uncertainty about their values so they claim that in the future they’ll have a better idea, just to put off having to answer a WTP question.

Kolstad and Guzman (1999) present a bidding model to show that if bidders are uninformed about their values and can acquire costly information about those values, then in a first price auction bidders will tend to overstate WTA and understate WTP. The information acquisition is different from Zhao and Kling, since individuals have the option to purchase information before the auction. The Kolstad and Guzman model has not been estimated or tested against alternatives.

The idea that uncertainty over one’s values can explain patterns in nonmarket valuation has been advanced in several papers. These explanations are not necessarily focused on WTA/WTP experiments.

3.4.2 Experimental Effects

A second line of argument is that the WTA/WTP disparity is an experimental artifact that disappears in real world economic contexts. Therefore, no defense of the neoclassical model is necessary.

Experience

The idea that the lack of experience was responsible for the WTA-WTP gap was first explored by Coursey, Hovis and Schulze (CHS). They were responding in particular to the Knetsch and Sinden findings of disparities in WTA and WTP for lotteries. CHS argued, as have

others that followed, that the mechanism employed by Knetsch and Sinden did not allow for learning. Providing subjects the opportunity to learn how the procedures work has been common practice in experimental economics.²⁰ That learning about the elicitation procedure might be valuable procedures seems quite reasonable for subtle mechanisms such as the Vickrey second price auction or the Becker-deGroot-Marschak mechanism. One might argue that the Knetsch-Sinden mechanism was quite transparent—accept a payment of \$2 to give up a lottery for subjects who received one color ticket or pay \$2 to play the same lottery for those who received another color ticket. Note that the literature has often confounded the idea that individuals need to learn about the mechanism (auction, for example) with the idea that they need to learn about the good being valued (lottery, for example).

The CHS experiments allowed learning within the study in several ways. Using a modification of a Vickrey auction, CHS conducted three experiments. The first experiment was strictly hypothetical and therefore allowed learning about the elicitation procedure. The second experiment, consisting of a trial and a follow-up, allowed the respondents to experience the good—a small swallow of a bitter tasting substance—that they were paying to avoid or accepting payment to taste. The third experiment was a set of auctions conducted until an equilibrium allocating four cups of the bitter substance among eight subjects was reached.

Learning is quite evident in the CHS experiments. Mean WTA across all subjects started at about \$9.60 and mean WTP about \$2.50, significantly and substantially different. At the ending trial, WTA had fallen substantially to about \$4.50 and WTP had risen slightly to about \$3. The two values were not statistically different at the last trial.²¹ The study supported the idea that learning reduces the WTA-WTP gap but it was hampered by a small number of observations.

The strongest evidence for a falling ratio comes from SSHK. They carried out a sequence of Vickrey auctions, with one of the auctions randomly chosen being played for real; that is, the individuals either paid (WTP) or were paid (WTA) the appropriate dollar sum and receiving either the more or less contaminated sandwich, if they were a winner of that auction round. They found that the WTA/WTP ratio fell substantially between the first and middle auction rounds.

Horowitz and McConnell (2002) reviewed the available studies as of 2002. They concluded that the evidence that the ratio would fall with experience was weak; a few studies

showed substantial drops but others did not. They further note that (i) even in studies where the ratio dropped, it dropped to levels that still were quite high (as judged, say, by equation (7)) and (ii) real world circumstances in which the WTA/WTP disparity might be important often would not have opportunities for individuals to learn. The “one shot” treatments that were used in many experiments are characteristic of many real world decisions.

Plott and Zeiler (2005) investigated the impact of learning on WTA-WTP gap with particular emphasis on the experimental procedures. They found that when they implemented an extensive set of controls (“incentive-compatible elicitation device, training, paid practice, and anonymity”) there was no statistical difference between WTA and WTP. They concluded that paid practice rounds were unnecessary because training was sufficiently effective.

The learning and experience referenced so far took place in the lab. In contrast, List (2003) investigated experience as one might understand it intuitively. How would subjects who have a longer life history of transacting and trading (“dealers”) compare with subjects without such experience (non-dealers)? List provides evidence along these lines from several field experiments. In the first experiment, he found that when subjects were endowed with two similar-value sports cards, dealers traded about 45 percent of the time, while non-dealers, those with less experience, traded 20 to 25 percent of the time. Low trading ratios are consistent with but not the same as high WTA/WTP ratios.

The List finding is consistent with two possible models. In each case the inexperienced traders have an instantaneous endowment effect from the presence of one of the cards. The greater propensity of traders to exchange cards can be explained by the absence of an endowment effect *or* by the Randall-Stoll result that WTA and WTP are both equal to market price when there is a large market with low transactions costs. In other words, as with the pattern of WTA/WTP across goods, both the neoclassical model and the alternative reference-dependent model have similar qualitative predictions.

Survey Procedures

It is natural to examine whether other survey procedures, beyond subject experience and learning, are responsible for high WTA/WTP ratios. Horowitz and McConnell (2002), studying roughly 200 experiments in 50 academic articles, argued against the claim that survey effects were responsible. They cite three pieces of evidence: (i) Experiments involving real goods did

not have significantly lower ratios than experiments involving hypothetical goods. (ii) Experiments that used incentive compatible valuation methods yielded significantly higher ratios rather than lower. (iii) Student subjects had statistically significantly lower ratios than non-student subjects, even when accounting for survey procedure. It was important in this latter case to control for survey procedure because students did not participate in a random set of survey procedures.

Plott and Zeiler (2005) provides the most systematic effort to date to demonstrate that the WTA/WTP gap could be reduced or eliminated by appropriate, rigorous survey procedures.

3.5 Non-Neoclassical Models

Research into the WTA/WTP gap has pitted the behavioral ideas of loss aversion and the endowment effect against neoclassical theory. The structure of the behavioral ideas has recently been significantly tightened by the model of Köszegi and Rabin. In the longer run it seems reasonable that this model will be further refined and strengthened and a variety of forces will be found to contribute to the difference between WTA and WTP. We look at what seem to be the more important determinants of the gap that relate significant influences on economic choices rather theories of preferences.

The Köszegi and Rabin model of reference dependent preferences is useful in helping to understand when to expect WTA/WTP gaps as well as the absence of gaps. The model, with its construction of the expected reference point, conceptually divorces the status quo from the reference point when appropriate. Hence experienced traders in the List model would not expect to retain the sports card with which they were endowed. Likewise in the Plott and Zeiler (2007) experiments, the experimental procedures effectively separated the endowment from the reference point with careful instructions.

The question about the relevance of different types of experience remains. List's results show that experienced traders carry their knowledge to field experiments. Does the learning that takes place in a lab persist between lab experiments? Does learning in one type of elicitation such as BDM carry over to another type, such as a Vickrey auction? And finally, if we are interested in the persistence of an implicit WTA/WTP gap in real world transactions, can we expect WTA/WTP gaps for rare but significant contracts such as housing purchases or marriage, where there is little opportunity to learn?

3.6 Concluding WTA/WTP comments

Another way to view the WTA/WTP literature is to recognize that if WTA/WTP “fails” then it must be because either WTA or WTP, or both, have been mis-measured. The NOAA Blue Ribbon Panel implicitly argued that WTA was the culprit. The consequences of this literature *for estimation of WTA* have not been much explored.

4. Consistency, Preference Reversal, and Related Tests

This section looks at a set of experiments that have often been confused with WTA/WTP experiments but are instead more powerful and more damning of the neoclassical model. That they have frequently been lumped together with WTA/WTP experiments is a bit surprising, although the confusion shows how appealing the WTA/WTP framework is. We label these *consistency experiments*.

Consistency in this context means that an individual who weakly prefers A to B must not strictly prefer B to A.²² Consistency is such a fundamental notion to the idea of rationality that it is difficult to imagine the social sciences without it, although it is not difficult to set up experiments in which it fails. Introductions to utility theory claim that rational preferences must be complete, transitive, and reflexive. Consistency is a version of transitivity.

A small number of experiments have tested consistency directly and almost all have shown that it to fail. Given the key role of consistency in standard economics it is odd that its violations have been overlooked. In large part, this failure is because the experiments have been lumped in with WTA/WTP experiments. None of these consistency or preference reversal experiments has been conducted with environmental goods to our knowledge.

4.1 Preference Consistency

Early experimenters treated preference reversal (for choice under uncertainty), non-exponential discounting (for choice over time), and high WTA/WTP ratios as evidence of inconsistent preferences.²³ In each of these cases, subsequent analysis showed that such behaviors do not directly violate consistency,²⁴ although they strongly hint at it. Direct tests of consistency have been less common.

Knetsch (1989) is the first example we can find of a direct experimental test of consistency for ordinary goods not involving time or uncertainty. In one treatment he gave student subjects a mug and told them they could exchange it for a bar of chocolate. In another treatment he gave students a bar of chocolate and told them they could exchange it for a mug. If choices are consistent, the proportional of individuals choosing each item should be the same regardless of which item they were first handed. Instead, Knetsch found that 89 percent of the individuals chose the mug when handed the mug to start with while just 10 percent chose the mug when handed the chocolate to start with. If we assume that individuals in the two treatments have similar preferences then these experiments provide strong evidence against consistent preferences.

Kahneman, Knetsch, and Thaler (1990) conducted a number of similar experiments using choices between money and a good rather than between two goods. Although KKT couched their experiments in terms of WTA and WTP, the tests also provide a test of consistency. They again found evidence against.

To see the difference between WTA/WTP and the consistency experiments, consider the following set-up. In treatment one, a group of individuals is given initial endowment q_1 . They have the opportunity to surrender it and receive a sum of money, x . They face the choice:

$$(8) \quad V(y, q_1) \text{ vs. } V(y+x, q_0)$$

Note that if x is above the individual's WTA then the individual should always choose the first bundle, whereas if x is below the individual's WTA she should choose the second. By varying the amount x across individuals, the experimenter is able to infer the distribution of WTA.

In treatment two, a second group of individuals is given an amount of money x and initial endowment q_0 . They have the opportunity to purchase q_1 for the price x . They face the choice:

$$(9a) \quad V(y+x, q_0) \text{ vs. } V(y, q_1)$$

Of course, the choices in (8) and (9a) are exactly the same and should elicit the same behavior.

To see the relationship to the WTA/WTP literature, replace $y+x$ by y' . Rewrite (9a) as:

$$(9b) \quad V(y', q_0) \text{ vs. } V(y'-x, q_1)$$

Suppose we assume that choices are insensitive to income, a reasonable and presumably innocuous assumption. Then if x is above the individual's WTP she should choose the first option and if x is below the individual's WTP she should choose the second option. By varying the amounts x it is possible to treat (9b) as eliciting WTP. Since experiments show that individuals are more likely to choose the first option in each treatment, the implied WTA is greater than the implied WTP.

The assumption that choices are insensitive to income is a perfectly reasonable empirical assumption but it keeps these experiments from providing true tests of WTA vs. WTP since the sensitivity of choices to income is the crucial feature of the WTA-WTP relationship.²⁵ On the other hand, the KKT experiments (along with Bateman et al. 1997 and Morrison) provide tests of the much more crucial property of consistency. It is worth noting that KKT and similar experiments could also be used to show that WTA/WTP ratios violate neoclassical predictions if the analysis were to include participant income as an explanatory variable.

More recent tests have dropped the reference to WTA-WTP and focused more explicitly on consistency. Horowitz, List, and McConnell (2008) report a direct test based on experiments conducted by John List. In one treatment he gave subjects a single sports card of type A, which they could freely trade for one of type B. In a second treatment he gave subjects a sports card of type B and allowed them to freely trade it for type A. Let the percent of subjects who opted to trade A for B be p . If choices are consistent, and assuming that individuals in the two treatments have similar preferences, then the percent of subjects who opted to trade B for A in the second treatment should be $100-p$. In the first treatment, 22.5 percent of the individuals opted for a trade; in the second treatment, 27.5 percent of individuals opted for trade. The percentage of subjects who opted for card B in treatment two, 72.5, was significantly different from the percent who opted for B in treatment one. Various repetitions of this and similar choice experiments showed a robust pattern of inconsistent preferences.

A second set of experiments showed similar behavior when collective choice was involved, a situation that is particularly relevant for environmental decision-making. In a set of treatments in which individuals voted whether to trade, with all individuals having to make the trade if more than 50 percent voted in favor, he found the same pattern: The proportion of

individuals voting in favor of making a trade in the direction of B was significantly different from the individuals voting in favor of not making a trade in the direction of A.

These consistency tests have the shortcoming that they have not been conducted either with higher value items (goods vs. money tests, as in Kahneman, Knetsch, and Thaler) or with items that have a greater discrepancy in value (goods vs. goods tests, as in Knetsch or Horowitz, List, and McConnell). Experiments using goods with a greater discrepancy in values would be useful since one could argue that when the two choice goods are similar in value individuals are close to indifferent between them and therefore choices do not yield much information.

The consistency tests are more powerful than tests of WTA-WTP for two reasons. First, they require no assumptions about income effects. There is no unobservable feature or component, however minor, that confounds the comparison of (8) and (9a). Second, no statistical procedure is needed to infer WTA or WTP from choices at different levels of x .

4.2 Preference Reversal

It is worth briefly covering preference reversal experiments since they played an important early role in experimental economics and are close, although not identical, to consistency experiments.

A narrow definition of preference reversal is that an individual prefers A to B but is willing to pay more for B than for A. This phenomenon was first observed for lotteries. Grether and Plott showed that such patterns could be elicited in choices over lotteries. In an important paper, Karni and Safra showed that this sort of preference reversal could be due to preferences that were not linear in the probabilities, also known as non-expected utility, and therefore that preference reversal was not a manifestation of inconsistent preferences.

Although preference reversal experiments for lotteries are not technically able to uncover preference inconsistency they provide evidence that could lead to contradictory public policy recommendations. In the public health arena, Ryan and San Miguel (2000) examined preferences over two alternative treatments for a condition known as menorrhagia. If treatment A were preferred over treatment B then individuals should be willing to pay more for A than for B, they argue. They showed that 30 percent of their subjects failed this test. No such test has been performed for environmental goods to our knowledge.

Several other forms of experimental behavior have been termed preference reversal that do not fit this precise framework. The most prominent of these is the lives-saved versus lives-lost example of Kahneman and Tversky, an example so well known that it has its own Wikipedia entry. The article that contained this example, along with other examples of behavioral anomalies, is one of the most cited social science papers.

The lives-saved versus lives-lost experiment is a “true” test of inconsistent preferences (rather than an indirect test, as preference reversal experiments are) but because the options contain subtle changes in wording, this experiment has fallen into the framing literature. It is worth noting that the Kahneman-Tversky example relies on uncertainty, since one of the options contains uncertainty about the outcome of one of the life-saving programs. A similar example involving the environment and not involving uncertainty has not been found to our knowledge, although we suspect that this kind of inconsistent behavior could indeed be elicited.

5. Concluding Comments

The neoclassical model is the underpinning of valuation and benefit-cost analysis. A vast number of experiments, of which those based on willingness-to-accept and willingness-to-pay are part, have now documented the many ways in which the neoclassical model fails to predict behavior. Researchers need to address how those failures affect both specific valuation techniques and the entire valuation enterprise.

In many fields of economics the implications of neoclassical failure are not dire. For contracts, incentives, auctions, and related institutions that have been much studied by experimentalists, behavioral findings have clever lessons that can largely be applied without raising questions about the underlying purposes of contracts or auctions. For environmental valuation, which has a much larger normative component, the neoclassical model’s failure is more central since it calls into question the underlying normative motivation for valuation. The neoclassical model is both normative and positive; in the case of valuation, a failure on the positive side also weakens normative conclusions.

The neoclassical model underlies valuation in two ways. First, nonmarket valuation requires that individual choices be “consistent.” Second, the neoclassical framework provides a standard, parsimonious set of variables that “count” for welfare: prices, income, available substitutes, timing. When the neoclassical model fails it is because nonmarket values are

sensitive to some variable that we otherwise would treat as irrelevant to welfare, such as reference points, or are insensitive to variables that we would treat as key, such as the amount of the good being provided.²⁶ This parsimonious and widely adopted notion of welfare has been the key to the flourishing of economics.

When choices are sensitive to these “extraneous” variables, economists *could* continue conduct nonmarket valuation studies, provided choices passed the test of consistency. But it would hard, if not impossible, to incorporate these variables in welfare calculations. The problem is not necessarily in including these variables in the valuation exercise – in most cases, it is possible to design valuation experiments that include almost any mixture of policy elements, economic elements and, in particular, non-economic elements. The problem is in deciding what level of these non-economic variables is correct or desirable when counting welfare.

References

- Aadland, D. M., Caplan, A. J. and Phillips, O. R., 2007. A Bayesian examination of information and uncertainty in contingent valuation. *Journal of Risk and Uncertainty*, 35(2), 149-178.
- Aadland, David and Caplan, Arthur J., 2003. Willingness to Pay for Curbside Recycling with Detection and Mitigation of Hypothetical Bias. *American Journal of Agricultural Economics*, 85(2), 492-502.
- Aadland, David and Caplan, Arthur J., 2006. Cheap talk reconsidered: New evidence from CVM. *Journal of Economic Behavior and Organization*, 60, 562-578.
- Ajzen, Icek, Brown, Thomas C. and Carvajal, Franklin, 2004. Explaining the Discrepancy Between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation. *Personality and Social Psychology Bulletin*, 30(9), 1108-1121.
- Amiran, Edoh and Daniel Hagen, 2003. 'Willingness to pay and willingness and to accept: How much can they differ? Comment', *American Economic Review*, 93, 458-463.
- Arrow, Kenneth, Solow, Robert, Portney, Paul, Leamer, Edward E., Radner, Roy et al., 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58(10), 4602-4614.
- Bateman I., Burgess D, Hutchinson WG, et al. 2008. *Journal of Environmental Economics and Management*, 'Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness' 55, 127-141,
- Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R., 1997. A theory of reference dependent preferences. *Quarterly Journal of Economics* 112, 479-505.
- Bishop, R.C. and Heberlein, T. A., 1986, Does Contingent Valuation Work? In: R. Cummings, D. Brookshire and W. Schulze (Eds.), *Valuing Environmental Goods: A State of the Art Assessment of the Contingent Valuation Method*. Rowman and Allenheld, Totowa, NJ, pp. 123-147.
- Bishop, Richard C. and Heberlein, T.A., 1979. Measuring Values of Extramarket Goods: Are Indirect Measures Biased? *American Journal of Agricultural Economics*, 61(5), 926-930.
- Blumenschein, K., Blomquist, G. C., Johannesson, M., Horn, N. and Freeman, P. R., 2008. Eliciting Willingness to Pay without Bias: Evidence from a Field Experiment. *Economic Journal*, 118(525), 114-137.
- Blumenschein, K., M. Johannesson, Blomquist, G. C., B. Liljas, and R.M. O'Connor. 1998. 'Experimental results on expressed certainty and hypothetical bias in contingent valuation', 65, 169-177.
- Bockstael, N.E. and K.E. McConnell 2006. *Environmental and Resource Valuation with Revealed Preferences*. Springer: Dordrecht, The Netherlands.
- Bohm, Peter, 1972. Estimating the Demand for Public Goods: An Experiment. *European Economic Review*, 3, 111-130.
- Boulier, Bryan and Robert Goldfarb. 1998 'On the use and nonuse of surveys in economics', *Journal of Economic Methodology*, 5(1), 1-21.
- Brookshire, D. S. and Coursey, D. L., 1987. Measuring the Value of a Public Good: An Empirical Comparison of Elicitation Procedures. *American Economic Review*, 77(4), 554-566.
- Brown, T. C., Champ, P., Bishop, R. and McCollum, D., 1996. Which Response Format Reveals the Truth About Donations to a Public Good? *Land Economics*, 72(2), 152-166.

- Brown, Thomas C., Ajzen, Icek and Hrubes, Daniel, 2003. Further Tests of Entreaties to Avoid Hypothetical Bias in Referendum Contingent Valuation. *Journal of Environmental Economics and Management*, 46(2), 353-361.
- Bulte, Erwin, Gerking, Shelby, List, John A. and Zeeuw, Aart de, 2005. The effect of varying the causes of environmental problems on stated WTP values: evidence from a field study. *Journal of Environmental Economics and Management*, 49, 330-342.
- Burton, Anthony C., Carson, Katherine S., Chilton, Susan M. and Hutchinson, W. George, 2003. An Experimental Investigation of Explanations for Inconsistencies in Responses to Second Offers in Double Referenda. *Journal of Environmental Economics and Management*, 46(3), 472-489.
- Carson, Richard T. and Groves, Theodore, 2007. Incentive and informational properties of preference questions. *Environmental and Resource Economics*, 37, 181-210.
- Carson, Richard T., Mitchell, Robert C., Hanemann, W. Michael, Kopp, Raymond J., Presser, Stanley et al., 1992. A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill. A Report to the Attorney General of the State of Alaska.
- Champ, Patricia, Richard Bishop, Thomas Brown and Daniel McCollum, 1997. "Using donation mechanisms to value nonuse benefits from public goods", *Journal of Environmental Economics and Management*, **33**,151-162.
- Chattopadhyay, S. 2002. 'Divergence in alternative Hicksian welfare measures: the case of revealed preference for public amenities', *J. of Applied Econometrics*, 17: 641-666.
- Corrigan, J.R., C.L. Kling and J.-H. Zhao, 2008. Willingness to Pay and the Cost of Commitment: An Empirical Specification and Test', *Environmental and Resource Economics*, **40**, 285-298.
- Coursey, D. L., Hovis, J. L. and Schulze, W. D., 1987. The Disparity between Willingness to Accept and Willingness to Pay Measures of Value. *Quarterly Journal of Economics*, 102(3), 679-690.
- Cummings, R. G., Harrison, G. W. and Rutström, E. E., 1995a. Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible? *American Economic Review*, 85(1), 260-266.
- Cummings, R. G. and Taylor, L. O., 1999. Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method. *American Economic Review*, 89(3), 649 - 665.
- Cummings, R.G., Brookshire, D.S and Schulze, W.D. (Eds.), 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Rowman and Allanheld, Totowa, NJ.
- Cummings, R.G., Harrison, G. W. and Osborne, L.L., 1995b. Can the Bias of Contingent Valuation Surveys Be Reduced? B-95-03, Division of Research, College of Business Administration, Univ. of South Carolina, Columbia, SC.
- Cummings, Ronald G., Elliott, Steven, Harrison, Glenn W. and Murphy, James, 1997. Are Hypothetical Referenda Incentive Compatible? *Journal of Political Economy*, 105(3), 609-621.
- Cummings, Ronald G. and Taylor, Laura Osborne, 1998. Does Realism Matter in Contingent Valuation Surveys? *Land Economics*, 74(2), 203-215.
- Diamond, Peter A. and Hausman, Jerry A., 1994. Contingent Valuation: Is Some Number Better Than No Number? *Journal of Economic Perspectives*, 8(4), 45-64.

- Dickie, M., Fisher, A. and Gerking, S., 1987. Market Transactions and Hypothetical Demand Data: A Comparative Study. *Journal of the American Statistical Association*, 82, 69-75.
- Dillman, Don A., 1978. *Mail and Telephone Surveys: The Total Design Method*. Wiley, New York.
- Friedman, Milton. 1953. 'The Methodology of Positive Economics' in *Essays in Positive Economics*, The University of Chicago Press, Chicago.
- Haab, Timothy C., Ju-Chin Huang and John C. Whitehead, 1999. 'Are hypothetical referenda incentive-compatible? A comment', *Journal of Political Economy*. **107**,86-96.
- Hammack, J. and G. M. Brown. 1974. *Waterfowl and Wetlands: Towards Bioeconomic Analysis*, Baltimore, Md.: Johns Hopkins Press.
- Hanemann, W. Michael, 1984. Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses. *American Journal of Agricultural Economics*, 66(3), 332-341.
- Hanemann, W. Michael, 1991. 'Willingness to pay and willingness to accept: How much can they differ?', *American Economic Review*, **81**, 635-647.
- Hanemann, W. Michael, 1994. Valuing the Environment Through Contingent Valuation. *Journal of Economic Perspectives*, 8(4), 19-43.
- Harrison, Glenn W., 2002. *Experimental Economics and Contingent Valuation*, University of South Carolina, Columbia, SC.
- Harrison, G. W., 2006. Experimental evidence on alternative environmental valuation methods. *Environmental and Resource Economics*, 34(1), 125-162.
- Harrison, G. W., Harstad, R.M. and Rutström, E.E., 2004. Experimental Methods and Elicitation of Values. *Experimental Economics*, **7**, 123-140.
- Harrison, G.W. and Rutström, E.E., forthcoming, Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods. In: C. Plott and V.L. Smith (Eds.), *Handbook of Results in Experimental Economics*. Elsevier Science, New York.
- Hausman, Jerry A., ed, 1993. *Contingent Valuation: A Critical Assessment*. Contributions to Economic Analysis. Amsterdam: North Holland Publishers.
- Heberlein, T. A. and Bishop, R., 1986. Assessing the Validity of Contingent Valuations: Three Field Experiments. *Science of the Total Environment*, 56, 434-479.
- Hoehn, J.P. and Randall, A., 1987. A Satisfactory Benefit Cost Indicator from Contingent Valuation. *Journal of Environmental Economics and Management*, 14(3), 226-247.
- Horowitz, John, John List and K.E. McConnell, 2007. 'A Test of Diminishing Marginal Value', *Economica*, **74**, 650-663.
- Horowitz, J. and K.E. McConnell. 2002 "A Review of WTA/WTP Studies" *Journal of Environmental Economics and Management* **44** 426-447.
- Horowitz, J. and K.E. McConnell 2003. "Willingness to Pay, Willingness to Accept and the Income Effect", *Journal of Economic Behavior and Organization* **51**, 537-545.
- Kealy, M., Dovidio, J. and Rockel, M., 1988. Accuracy in Valuation is a Matter of Degree. *Land Economics*, 64, 158-171.
- Knetsch, Jack, 1989. 'The endowment effect and evidence of irreversible indifference curves', *American Economic Review*, **79**, 1277-1284.
- _____. 1995. "Assumptions, Behavioral Findings, and Policy Analysis", *J of Policy analysis and Management*, **14**, 68-78.
- Kolstad, Charles, and Rolando Guzman, 1999. "Information and the Divergence between Willingness to Accept and Willingness to Pay," *JEEM*, **38**, 66-80.

- Köszegei, Botond, and Matthew Rabin, 2006. "A Model of Reference-Dependent Preferences", *Quarterly Journal of Economics* **121**:1133-1165.
- Krutilla, John. 1967. 'Conservation Reconsidered', *American Economic Review*, 57(4): 777-786.
- Landry, C. E. and List, J. A., 2007. Using ex ante approaches to obtain credible signals for value in contingent markets: Evidence from the field. *American Journal of Agricultural Economics*, 89(2), 420-429.
- LaPiere, R.T., 1934. Attitudes vs. Actions. *Social Forces*, 13, 230-237.
- List, J. A., 2001. Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards. *American Economic Review*, 91(5), 1498-1507.
- List, J.A. 2003. 'Does Market Experience Eliminate Market Anomalies?', *Quarterly Journal of Economics*, **118**, 41-71
- List, J.A. 2004. 'Neoclassical Theory versus Prospectus Theory: Evidence from the Market Place', *Econometrica*, **72**, 615-625
- List, J.A. and Shogren, Jason, 1998. 'Calibration of the difference between actual and hypothetical valuations in a field experiment', *Journal of Economic Behavior and Organization*, **37**, 193-205.
- List, J. A. and Shogren, J. F., 2002. Calibration of Willingness to Accept. *Journal of Environmental Economics and Management*, 43(2), 219-233.
- List, John A. and Gallet, Craig, 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics*, 20, 241-254.
- Loomis, J., Brown, T., Lucero, B. and Peterson, G., 1996. Improving Validity Experiments of Contingent Valuation Methods: Results of Efforts to Reduce the Disparity of Hypothetical and Actual Willingness to Pay. *Land Economics*, 72(4), 4450-4461.
- Loomis, John, Gonzalez-Caban, Armando and Gregory, Robin, 1994. Do reminders of substitutes and budget constraints influence contingent valuation estimates? *Land Economics*, 70(4), 499-506.
- Lusk, J. L., 2003. Effects of cheap talk on consumer willingness-to-pay for golden rice. *American Journal of Agricultural Economics*, 85(4), 840-856.
- Mansfield, Carol. 1999. 'Despairing over Disparities: Explaining the Difference between Willingness to Accept and Willingness to Pay', *Environmental and Resource Economics*, **13**, 219-234.
- Mitchell, Robert Cameron and Carson, Richard T., 1986, Some Comments on the State of the Arts Assessment of the Contingent Valuation Method Draft Report. In: R.G. Cummings, D.S. Brookshire and W.D. Schulze (Eds.), *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Rowman and Allanheld, Totowa, NJ.
- Mitchell, Robert Cameron and Carson, Richard T., 1989, Using Surveys to Value Public Goods: The Contingent Valuation Method. Resources for the Future, Washington, DC.
- Morrison, Gwendolyn, 1997, 'Willingness to pay and willingness to accept: some evidence of an endowment effect', *Applied Economics*, **29**, 411-417.
- Murphy, James J., Stevens, Thomas H., Allen, P. Geoffrey and Weatherhead, Darryl, 2005a. A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics*, 30(3), 313-325.

- Murphy, James J., Stevens, Thomas H. and Weatherhead, Darryl, 2005b. Is Cheap Talk Effective at Eliminating Hypothetical Bias in a Provision Point Mechanism? *Environmental and Resource Economics*, 30(3), 313-325.
- Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W. and McGuckin, T., 1994. Hypothetical Surveys and Real Economic Commitments. *Land Economics*, 70(2), 145-154.
- Plott, Charles R. and Kathryn Zeiler, 2005. 'The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Values', *American Economic Review*, **97**, 530-545.
- Plott, Charles R. and Kathryn Zeiler, 2007. 'Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory?', *American Economic Review*, **97**, 1449-1466
- Poe, G.L., Clark, J.E., Rondeau, D. and Schulze, W.D., 2002. Provision Point Mechanisms and Field Validity Tests of Contingent Valuation. *Environmental and Resource Economics*, 23, 105-131.
- Polomme, P., 2003. Experimental Evidence on Deliberate Misrepresentation in Referendum Contingent Valuation. *Journal of Economic Behavior and Organization*, 52(3), 387-401.
- Portney, Paul, 1994. The Contingent Valuation Debate: Why Should Economists Care? *Journal of Economic Perspectives*, 8(4), 3-17.
- Randall, Alan and John Stoll, 1980. 'Consumer's Surplus in Commodity Space', *American Economic Review*, **70**, 449-455.
- Rowe, R. and L. Chestnut, editors. 1983. *Managing Air Quality and Visual Resources at National Parks and Wilderness Areas*, Westview Press.
- Schuman, Howard and Johnson, Michael P., 1976. Attitudes and Behavior. *Annual Review of Sociology*, 2, 161-207.
- Shogren, Jason F., 2005, Experimental Methods and Valuation in: K.G. Mäler and J. Vicent (Eds.), *Handbook of Environmental Economics*. Volume 2, Valuing Environmental Changes. North Holland, Amsterdam, pp. 969-1027.
- Shogren, Jason, Seung Shin, Dermot Hayes and James Kliebenstein, 1994. 'Resolving Difference in Willingness to Pay and Willingness to Accept', *American Journal of Agricultural Economics*, **84**, 270.
- Sinden, J. A., 1988. Empirical Tests of Hypothetical Biases in Consumers' Surplus Surveys. *Australian Journal of Agricultural Economics*, 32(2&3), 98-112.
- Smith, V.K. 1999. 'Of Birds and books: more on hypothetical referenda', *Journal of Political Economy*, **102**, 197-200.
- Smith, V. K. and Mansfield, C., 1998. Buying Time: Real and Hypothetical Offers. *Journal of Environmental Economics and Management*, 36, 209-224.
- Smith, Vernon L., 1976. Experimental Economics: Induced Value Theory. *American Economic Review*, 66(2), 274-279.
- Sugden, R., 1999. Alternatives to the neoclassical theory of choice. Chapter 6. In: I. Bateman and K.G. Willis (Eds.). *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EU, and Developing Countries*, Oxford University Press, 152-180.
- Taylor, L., 1998. Incentive Compatible Referenda and the Valuation of Environmental Goods. *Agricultural and Resource Economics Review*, 27(2), 132-139.

- Vossler, C. A. and McKee, M., 2006. Induced-value tests of contingent valuation elicitation mechanisms. *Environmental and Resource Economics*, **35**(2), 137-168.
- Whitehead, John C. and Cherry, Todd L., 2007. Willingness to pay for a Green Energy program: A comparison of ex-ante and ex-post hypothetical bias mitigation approaches. *Resource and Energy Economics*, **29**(4), 247-261.
- Willig, Robert, 1976, 'Consumer's surplus without apology', *Amer. Econ. Rev.* **66**, 589-597.
- Zhao, Jinhua, and Catherine Kling, "A new explanation for the WTP/WTA disparity," *Economics Letters* (2001) 73: 293-300.

¹The earliest explicit model of valuation was Hotelling's 1947 letter outlining the travel cost model.

²Ridker and Henning estimated effect of air pollution on housing values in St Louis.

³The controversy surrounding economists' use of interview data is given nicely in Boulier and Goldfarb. The engagement began with the so-called 'Richard Lester-Fritz Machlup' debate in which Lester concluded from interviewing businessmen that they did not behave according to marginal cost pricing and Machlup argued essentially that such interview data were unreliable. Milton Friedman helped dispense with the need for interviewing businessmen, at least for insight into business decisions, by arguing that competition forced businesses to behave according to marginal cost pricing (p. 22), making interview data superfluous. Of course, there is no equivalent mechanism to ensure that consumers allocate their budgets efficiently.

⁴Krutilla clearly stated the idea by noting, for example, that the existence of a fragile ecosystem is part of the real income of many individuals but does not contribute to the area under the demand curve for the resource. This came to be called existence value later.

⁵ Exceptions include: Taylor 1998; Burton *et al.* 2003; Polomme 2003; Vossler and McKee 2006.

⁶ A report to the Attorney General of Alaska (Carson *et al.* 1992) estimated that lost passive use values (also referred to as nonuse values) resulting from the spill were no less than \$2.8 billion.

⁷ The chapters in the book edited by Hausman 1993 were directed at weaknesses in contingent valuation. See also Diamond and Hausman.

⁸ There was some debate about whether the attitude-behavior literature in psychology is of relevance to CV. The controversy largely centered on a discussion of whether CV responses were expressions of attitudes, rather than intentions. Bishop and Heberlein 1986 make the case that this literature could provide a useful framework for understanding CV responses. After presenting some initial skepticism, Cummings *et al.* 1986 agree that the Bishop and Heberlein's argument has merits.

⁹ Carson and Groves 2007 suggest that this can be accomplished through careful survey design and judicious choice of words to develop a demand-revealing instrument that will induce people to truthfully respond in a way that maximizes their expected value.

¹⁰ Many field CV studies compare different hypothetical instruments, but do not include the real payment condition necessary to make any claims about the effects these have on hypothetical bias.

¹¹ They also had a hypothetical WTP treatment. However, there was not a corresponding real WTP treatment, and it would not be appropriate to compare the hypothetical WTP results with the real WTA results.

¹² Bishop and Heberlein 1986 disagree with Mitchell and Carson's conclusions.

¹³ They found no evidence of hypothetical bias with a fourth good (donations to a nonprofit organization to build and maintain bike trails) and therefore did not implement a cheap talk treatment.

¹⁴ Landry and List 2007 refer to these as consequential treatments to distinguish them from real payment treatments.

¹⁵ It is more precise to say that the good is "rationed at zero price."

¹⁶ Although Hammack and Brown was the most prominent early study of WTA and WTP, other environmental economists were investigating the issue around the same time, including Brown and Matthews (1970; salmon fishing) and Eby (1975; outdoor recreation). Jones-Lee also provided an early study (1976) on the value of life.

¹⁷ Hanemann's example is more involved than this but the lesson is the same. The Leontief framework has some odd properties, although these oddities disappear when preferences are smoother than Leontief. The analog to Hanemann's example is as follows. Suppose $p_h b^* < Y$. The individual receives so few buns that he does not spend all his income because he does not need more than b^* hamburgers. Next consider an increase in buns to b^*+d and assume $p_h(b^*+d) \leq Y$. That is, the individual continues not to spend all his income. Then willingness to pay for the increase d is $Y - p_h(b^*+d)$; that is, a non-negative amount. Willingness to accept a decrease, say from b^*+d to b^* ,

remains infinite. The positive willingness to pay in this circumstance is a bit misleading, however, because the individual was not spending this money to start with. Note that the willingness to pay for d is less than the unspent income in the initial situation, $Y - p_h b^*$.

¹⁸ AH state that “for each set of the public goods and for any fixed decrement in a public good, there exists an initial endowment of market goods so that no increment of the market goods than can compensate for the decrements in the public good” (p. 462). This statement is imprecise because consumers do not receive endowments of market goods. It is sufficient to restate the result as the individual having a specific income and facing specific prices such that the “initial endowment of market goods” is what he would purchase with that income at those prices.

¹⁹ This problem shows the third reason why the test in (6) is informative but ultimately unsatisfying, a combination of (i) the difficulty in the measuring income; (ii) the problem of determining whether the neoclassical model applies at the individual or household level; and (iii) the reliance of tests on measures of mean WTA, mean WTP, and mean $\frac{\partial WTP}{\partial y}$ when the theory applies at the individual level.

²⁰ In fact, it is often argued that the learning that takes place in a single bid with follow-up in discrete choice contingent valuation (double-bounded, DB) will lead to strategic behavior other than true-telling. See Hanemann, Cooper and Signorello. The idea that there are incentive effects in the DB CV has been challenged by Bateman et al. (2008), who show that learning takes place in a series of repeated DB CV experiments.

²¹ These WTA-WTP figures are approximate, taken from Figure II of CHS.

²² This version of consistency should not be confused with intertemporal consistency.

²³ In a move that will drive philosophers crazy, we do not distinguish between inconsistent choices and inconsistent preferences.

²⁴ For the intertemporal case, X showed non-exponential discounting was consistent with well-behaved preferences. For choice under uncertainty, Karni and Safra showed that preference reversal was consistent with well-behaved preferences.

²⁵ The sensitivity of revealed choices to income is small, consistent with the assumption maintained for many experiments.

²⁶ It is not clear who first made this observation about the twin problems of excessive sensitivity to items that should not matter and insensitivity to item that should.