Fall 2012

From Exxon to BP: Has Some Number Become Better Than No Number?

Catherine L. Kling
Iowa State University, ckling@iastate.edu

Daniel J. Phaneuf
University of Wisconsin–Madison

Jinhua Zhao
Michigan State University

From Exxon to BP: Has Some Number Become Better Than No Number?

Abstract
On March 23, 1989, the Exxon Valdez ran aground in Alaska's Prince William Sound and released over 250,000 barrels of crude oil, resulting in 1300 miles of oiled shoreline. The Exxon spill ignited a debate about the appropriate compensation for damages suffered, and among economists, a debate concerning the adequacy of methods to value public goods, particularly when the good in question has limited direct use, such as the pristine natural environment of the spill region. The efficacy of stated preference methods generally, and contingent valuation in particular, is no mere academic debate. Billions of dollars are at stake. An influential symposium appearing in this journal in 1994 provided arguments for and against the credibility of these methods, and an extensive research program published in academic journals has continued to this day. This paper assesses what occurred in this academic literature between the Exxon spill and the BP disaster. We will rely on theoretical developments, neoclassical and behavioral paradigms, empirical and experimental evidence, and a clearer elucidation of validity criteria to provide a framework for readers to ponder the question of the validity of contingent valuation and, more generally, stated preference methods.

Disciplines
Agricultural and Resource Economics | Economics | Environmental Law | Oil, Gas, and Energy

Comments
This article is from *Journal of Economic Perspectives* 26 (2012): 3–26, doi:10.1257/jep.26.4.3. Posted with permission.

Journal of Economic Perspectives—Volume 26, Number 4—Fall 2012—Pages 3–26

From Exxon to BP: Has Some Number Become Better than No Number?

Catherine L. Kling, Daniel J. Phaneuf, and Jinhua Zhao

On March 23, 1989, the *Exxon Valdez* ran aground in Alaska's Prince William Sound and released over 250,000 barrels of crude oil, resulting in 1300 miles of oiled shoreline, the deaths of 250,000 birds, 2800 otters, over 250 seals, and destruction of nearly uncountable salmon and herring eggs (for details, see http://www.evostc.state.ak.us/facts/index.cfm). This event and its aftermath, graphically illustrated to television viewers around the world, ignited debate about the environmental risks of oil usage, the adequacy of regulatory oversight, and the appropriate compensation for damages suffered. The Exxon spill also ignited a debate within the economics profession concerning the adequacy of methods to value public goods, particularly when the good in question has limited direct use, such as the pristine natural environment of the spill region.

Shortly following the *Valdez* grounding, as legal and regulatory processes began, representatives of the state of Alaska, the U.S. government, and Exxon sought expertise in valuing public goods for the purpose of measuring lost economic value from the spill. In turn, a therefore relatively obscure technique referred to as the *contingent valuation* method received considerable attention. In the contingent valuation method, standard measures of economic value such as willingness to pay or willingness to accept are estimated using responses to survey questions. In contemporary lingo, contingent valuation is part of a broader category of approaches known as *stated preference* methods, which rely on peoples' responses

■ *Catherine Kling is Professor of Economics, Iowa State University, Ames, Iowa. Daniel Phaneuf is Associate Professor of Agricultural and Applied Economics, University of Wisconsin, Madison, Wisconsin. Jinhua Zhao is Professor of Economics, Michigan State University, East Lansing, Michigan. Their email addresses are ckling@iastate.edu, dphaneuf @wisc.edu, and jzhao@msu.edu, respectively.*

http://dx.doi.org/10.1257/jep.26.4.3. doi=10.1257/jep.26.4.3

to questions about researcher-designed—and therefore hypothetical—changes in environmental quality.

The efficacy of stated preference methods generally, and contingent valuation in particular, is no mere academic debate. Billions of dollars are at stake. A contingent valuation study of the damages from the *Exxon Valdez* spill generated an estimate of $4.9 billion (Carson, Mitchell, Hanemann, Kopp, Presser, and Ruud 2003) in lost economic value. In contrast, a recreation demand study of the damages from the spill yielded an estimate of $3.8 million (Hausman, Leonard, and McFadden 1995). The key explanation for the thousand-fold difference is that the estimate from the contingent valuation study is associated almost entirely with *passive-use* or *non-use* value—the value that people place on something simply because it exists, even if they never directly use the good. In contrast, the recreation study only measured economic damages arising from the loss of actual visits to the area of the spill. The authors of the two Exxon studies acknowledged that their methodologies captured distinct values. Carson et al. pointed out that their survey of non-Alaskans meant that the values would be almost exclusively associated with passive use. Hausman, Leonard, and McFadden (p. 29) likewise wrote: "If $3.8 million seems low, the reader must recall that we have estimated only those damages associated with recreational use. Damages associated with commercial use or damages associated with so-called nonuse values are not included in our estimates." Ultimately the *Exxon Valdez* case was settled through a U.S. District Court consent decree in 1991 (*Exxon Valdez* Oil Spill Council, "Settlement") that has paid out approximately $1 billion in damages and over $2 billion in immediate responses and restoration efforts.

While the conceptual basis for passive use value has been clear since John Krutilla's (1967) contribution in the *American Economic Review*, the only available method for measuring it relies on stated preferences, which immediately raises questions for economists. Are stated preference estimates likely to be inaccurate and devoid of useful information, or can a well-constructed survey generate accurate predictions? After all, economists have long favored analysis that is based on what people do rather than what they say. Given the high stakes involved, stated preference methods came under intense scrutiny during the Exxon legal battle. In its wake, the National Oceanic and Atmospheric Administration (NOAA) in 1992 charged a "Blue Ribbon" panel with the task of studying the efficacy of the conginent valuation method (Arrow, Solow, Portney, Leamer, Radner, and Schuman 1993). An influential symposium appearing in this journal in 1994 subsequently provided arguments for and against the credibility of the method, and an extensive research program published in academic journals has continued to this day. The disparity between the estimates of passive use values and direct use values provide ample explanation for this scrutiny, but it is worth emphasizing that for pristine wilderness areas, passive use may be the largest component of value—and stated preference may be the only game in town when it comes to estimation. Thus, if stated preference approaches are deemed unreliable and environmental damage assessment is limited to direct impacts such as lost productivity, health effects, damaged fisheries, displaced recreation, and similar pathways, then the damage from oil spills, toxic releases, and other accidents in remote locations may result in comparatively small monetized losses.

On April 20, 2010, the *Deepwater Horizon* oil rig affiliated with BP suffered an explosion, triggering the release of nearly five million barrels of crude oil into the Gulf of Mexico—a spill 20 times as large as the *Exxon Valdez*. The accident again led to oiled beaches, the death of seabirds and marine wildlife, and the altering of poorly understood and complex ecosystems. As we write, economists and attorneys are at work drawing on existing studies and undertaking new ones to estimate the economic damages from the spill. Much of the work being conducted as part of the legal process is confidential and ongoing, though early evidence from a recreation study (Alvarez, Larkin, Whitehead, and Haab 2012) and a contingent valuation survey (Larkin 2012) is shortly to appear. BP has already set up a $20 billion trust fund for remediation of environmental damages, of which $6 billion was spent as of mid 2012 (Guarino 2012). The large amounts of money involved are once again likely to spur fundamental questions about the veracity of the public goods valuation methods available to economists. This time, however, two decades of research are on the table to guide the work and inform the debate.

The goal of this paper is to assess what occurred in the academic literature between the Exxon spill and the BP disaster in order to shed light on the fundamental question of the validity of contingent valuation and, more generally, stated preference methods. The two oil spills provide useful bookends for our discussion, and the drama surrounding them helps highlight the importance of public goods valuation for policy and litigation purposes. We stress, however, that the issue of stated preference efficacy is much broader than valuing damages from oil spills, and so most of the discussion that follows will be in a more general framework. In particular, we summarize the most salient findings from the now large stated preference literature.[1] The fundamental question is straightforward: are the values elicited from stated preference methods reliable enough to use in policy analysis and/or litigation? We will rely on theoretical developments, neoclassical and behavioral paradigms, empirical and experimental evidence, and a clearer elucidation of validity criteria to provide a framework for readers to ponder this question. Before doing so, however, we first provide a bit of history and then some necessary background on stated preference methods.

Historical Perspective

A search on the Thomson Reuters Web of Science using "contingent valuation" as the topic returns only 49 journal articles as of 1989. These papers, along with important books by Cummings, Brookshire, and Schulze (1986) and Mitchell and

[1] We make no attempt to review thoroughly the now extensive stated preference literature. Several reviews trace the literature from the earliest published suggestion of the method (Ciriacy-Wantrap 1947) and its first implementation by Davis (1963), through the refinements and applications studied in the 1970s, 1980s, and 1990s. See for example Randall, Ives, and Eastman (1974), Cummings, Brookshire, and Schulze (1986), Mitchell and Carson (1989), Carson, Flores, and Meade (2001), Carson and Hanemann (2005), and Bennett (2011).

Carson (1989), comprised the bulk of the published literature at that time. Shortly after the *Exxon Valdez* spill, Congress passed the Oil Pollution Act of 1990, which specifically included lost passive use value as a compensable damage. Congress charged the NOAA with identifying methods to value these damages and, facing the fallout from the Exxon debate, NOAA commissioned a panel chaired by Kenneth Arrow and Robert Solow and charged it with answering a deceptively simple question: Is the contingent valuation method capable of providing estimates of lost nonuse values that are reliable enough to be used in natural resource damage assessments? (The panel was also asked to consider whether passive use should be part of damage assessment, but its affirmative answer has not generated the same attention as its other findings, and so we do not consider it further here.) In January 1993, after reviewing the available literature and accepting testimony from researchers, the NOAA Panel provided its answer (Arrow, Solow, Portney, Leamer, Radner, and Schuman 1993, p. 43):

> [W]e identify a number of stringent guidelines for the conduct of CV [contingent valuation] studies. . . . The Panel concludes that under those conditions (and others specified above), CV studies relay useful information. We think it is fair to describe such information as reliable by the standards that seem to be implicit in similar contexts, like market analysis for new and innovative products and the assessment of other damages normally allowed in court proceedings. . . . Thus, the Panel concludes that CV studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive-use values.

However, the panel left no doubt that its members had very strong reservations with the method and emphasized their concern about several potential biases and problems identified in the literature at that time. They also provided a set of guidelines that effectively established a list of best practices for the design and implementation of contingent valuation surveys.

In reaching their conclusions, the panel cited evidence from only two studies that compared contingent valuation estimates to elicited actual values for public goods and three studies that compared contingent valuation responses with elicited prices for private goods. Based in part on these early tests of the method's accuracy, the panel concluded "that hypothetical markets tend to overstate willingness to pay for private as well as public goods." In the 1994 symposium in this journal, NOAA panel member Paul Portney (1994) provided an introduction to the contingent valuation methodology and traced the key legal and policy developments up through the completion of the panel's report. In two additional papers, W. Michael Hanemann (1994) argued in favor of the method and Peter Diamond and Jerry Hausman (1994) argued against, with the latter authors raising the provocative question "is some number better than no number?"

With the luxury of hindsight, it is now clear that considerable work remained to be done—either to provide convincing evidence of the method's accuracy or its lack

thereof. First, a commonly accepted set of criteria on how to judge whether stated preference studies were adequate for a given task was missing from the vernacular. Second, it was apparent that more theoretical work was needed to understand if and when stated preference studies should be expected to provide unbiased assessments of the underlying economic values. Finally, much empirical work was needed to test the theory and methods in a wide variety of empirical settings.

The economics profession has risen to the challenge. In contrast to the small literature available at the time of the Exxon spill, by 2010 when the BP disaster occurred, at least 25 books and over 2,500 additional journal articles had been published on contingent valuation. This count likely understates the full collection in that newer types of stated preference studies, including choice experiments, may not be flagged under the "contingent valuation" search. In addition, Carson (2011) has amassed a bibliography of over 7,500 studies, which includes many works not published in the peer-reviewed literature.

Stated Preference Methods: A Short Primer

In this section, we frame our discussion of stated preference accuracy by placing it within the larger context of valuing public goods, also referred to as nonmarket valuation, and explaining a few basics on how it works.[2] Two general approaches are available. One makes use of private behavior in related markets to measure the economic value of a nonmarket good such as environmental quality. For example, data on how far people are willing to travel to reach an outdoor recreation destination of a given quality can be used to estimate the tradeoffs people make between money spent on travel and environmental quality at recreation sites. This type of approach is known as *revealed preference*. Hedonic analysis of housing markets is another common type of revealed preference approach routinely applied to environmental goods. Rather than indirectly inferring value from activity in related markets, *stated preference* approaches directly question individuals via surveys to obtain the information needed to value the nonmarket good. In both approaches, the objective is to measure economic value for a change in a nonmarket good by predicting respondents' willingness to pay, or willingness to accept, for the change. For an increase in environmental quality, willingness to pay (more formally, "compensating variation") is the most the individual would be willing to exchange to achieve the improvement. Likewise willingness to accept ("equivalent variation") is the least the individual would accept to forgo the improvement.

There are different types of stated preference approaches. The best-known, and the subject of the Exxon-era debates, is contingent valuation. In a contingent valuation survey, people are asked questions directly related to their willingness to

[2] For more complete treatments, see Champ, Boyle, and Brown (2003), Batemen et al. (2002), or Phaneuf and Requate (forthcoming). For a review of the large literature using stated preference method in health economics see de Bekker-Grob, Ryan, and Gerard (2012).

pay for a specific environmental program, commonly in the form of a yes/no answer to a posted price. A second type of stated preference approach is a choice experiment (Louiviere, Hensher, and Swait 2000; Kanninen 2007), in which a person is asked to consider an environmental commodity that is defined by several attributes. The respondent is presented with discrete options that represent different bundles of the attribute levels and asked to select a preferred alternative. A defining characteristic of choice experiments is that a respondent completes multiple choice tasks and selects from three or more options during each task. While contingent valuation and choice experiments share many design elements, the incentives they present to respondents can differ. At the risk of some confusion, we use stated preference and contingent valuation somewhat interchangeably in this essay, both for continuity with the earlier debates and for simplicity. We stress, however, that insights from choice experiments represent an increasingly important component of the literature.

Stated preference surveys typically share similar structures. To value a specific policy change that moves an environmental resource from one well-defined state to another, the survey needs to first describe the environmental good to respondents in a way that is understandable for a lay participant while remaining true to the underlying science. It then needs to communicate the existing level of environmental quality as well as the change being proposed and, finally, the specific policy intervention that will be used to bring about the change. After the commodity has been described, a survey will typically explain the constructed market and method of payment. A best practice for contingent valuation is to describe the market as a referendum in which the respondents are asked whether they would vote for or against the project in a public vote. Since the answer to a question of this type provides only an upper or lower bound on a respondent's value, statistical methods are used to translate this information into an estimate of the distribution of economic value in the population (Haab and McConnell 2002).

A critical part of the referendum question design is the posted price that the respondent is "offered" in considering whether to vote for the project. A careful experimental design is necessary for efficient estimation of mean willingness-to-pay estimates and large sample sizes are generally needed to achieve the desired precision. Other constructed market details include the conditions for provision (for example, whether a majority must vote in favor) and timing of the project. When presenting the posted price, the survey should also describe the method of payment, which can be coercive or voluntary. The former is usually preferred and includes, for example, changes in property tax rates, surcharges on utility bills, or generally assessed fees. The respondent completes the survey by reading the material describing the issue and then deciding, based on personal preferences and her budget constraint, whether to vote "yes" or "no."

Though it is relatively straightforward to describe the components of a contingent valuation study, actual implementation requires attention to many details. Current best practice for survey design involves the iterative use of focus groups, one-on-one interviews, and pre-testing to verify that the commodity description and

constructed market are appropriate for the purposes of the research. A premium is placed on a high level of specificity in the good being valued and the program being evaluated, since vague or abstract descriptions have been shown to lead to unreliable responses. Also, it is generally accepted that the exercise should seek to value the policy package broadly, rather than the change in the commodity narrowly, since context details should matter for how economic value arises. Finally, most surveys include questions designed to gauge how well respondents understood the material, the confidence they have in their responses, and the rationality of their answers.

To help readers who are unfamiliar with stated preference surveys better understand how such surveys are presented to respondents, we provide an abbreviated version of a contingent valuation question from a study published by Loomis, Kent, Strange, Fausch, and Covich (2000) and used as an example in Haab and McConnell's text (2002). The study concerned the valuation of a set of ecosystem services that would be generated by the purchase of water rights from landowners along the South Platte River in Colorado. Detailed information on the proposed plan's effects on wildlife habitat, erosion control, recreational opportunities, and water purification was provided to respondents. An in-person interviewer then asked respondents the following:

> If the majority of households vote in favor of the South Platte River restoration fund, the 45 miles of river would look like (*in-person interviewer points to a figure showing increased water quality and fish and wildlife*). If a majority votes against, these 45 miles of the South Platte River would remain as they are today, as illustrated by (*in-person interviewer points to a figure showing current management*). If the South Platte River restoration fund was on the ballot in the next election and it cost your household $B each month in a higher water bill, would you vote in favor or against?

The dollar amount $B was randomly filled in with one of twelve values ($1, 2, 3, 5, 8, 10, 12, 20, 30, 40, 50, 100). Based on survey responses from 100 local respondents, Loomis et al. estimated an average willingness to pay of $21 per month or over $250 annually per household for the proposal.

Lessons from Theory: When Should Stated Preference Estimates Match Real Payments?

Economists have long believed that observation of actual behavior in which people bear the consequences of their actions is the key to understanding their motives. In turn, this predisposition has given rise to an inclination to doubt the accuracy of answers provided in a survey context, particularly if it involves reporting more than a factual outcome. Recently, however, researchers have developed theories describing how people behave while answering surveys, given the time and cognitive

energy needed for the task. This is our point of departure for understanding what the necessary conditions are for survey answers to reflect real economic values.

The main theoretical tool has been mechanism design, applied to the problem of understanding when it is in a person's best interest to thoughtfully and truthfully report preferences in a stated preference exercise. In response to a critique by Cummings, Harrison, and Rutström (1995) that survey participants do not have the incentive to answer stated preference questions accurately, Carson and Groves (2007) argue that the necessary conditions for truthful reporting involve using an elicitation mechanism that discourages strategic responses, and fielding the survey in a way that encourages respondents to believe that the study's results could ultimately influence their well-being. These conditions are known as *incentive compatibility* and *consequentiality*, respectively. The need for incentive compatibility in eliciting responses from the public is not new: theorists have long known that departures from single-shot, binary, binding outcome choices provide an incentive for self-interested participants to depart from selection of their most preferred option. Indeed, these arguments are part of what led the NOAA Panel to recommend using a binary choice format for contingent valuation elicitation. The last two decades, however, have seen a much more complete investigation into the many nuanced ways that the design features of a stated preference survey can affect choices.

An important example relates to contributions to public goods. Economic theory predicts that, due to the incentive to free ride, a person's voluntary contribution to a collective good will be smaller than that person's true willingness to pay. However, this incentive can play out in a surprising form in hypothetical surveys. If the respondent believes the survey will be used to decide on the ultimate provision of a public good, that person will have incentive to report *more* than true willingness to pay in a voluntary elicitation in which payment is not binding, in order to influence provision so as to have the opportunity to free ride—and contribute less than stated—should the provision become a reality. For example, Champ, Bishop, Brown, and McCollum (1997) find in a field experiment that hypothetical willingness to donate is substantially larger than the donations they actually collected for a public good with mainly nonuse value, though they are not able to compare either to estimates of the true willingness to pay.

Concern about the role of consequentiality in stated preference survey research arose relatively recently (Carson and Groves 2007). Rather than assuming that respondents have the incentive to answer untruthfully (or truthfully), the consequentiality argument suggests that there are no predictable incentives for an inconsequential survey. Specifically, if the respondent has no reason to believe that her answers will influence an outcome that she cares about (either directly, or indirectly by how the survey results are used), there is no reason to expect that the respondent has dedicated effort to the process, and so the meaningfulness of that person's answers cannot be judged. In contrast, if the survey is consequential in the sense that the respondent thinks its conclusions may ultimately influence something that the respondent cares about, she will have incentive to devote effort. In this case, the truthfulness of the respondent's answers hinges on other factors related to the incentive compatibility

and other characteristics of the survey. Continuing the public good example from above, the person has incentive to bid more than her true willingness to pay in the hypothetical voluntary payment survey only if she believes such an act will influence the probability of the good being provided. Absent this condition, there is no prediction we can make about how the respondent will answer the survey question. Herriges, Kling, Liu, and Tobias (2010) show that estimates of economic value from people who received a consequentiality reminder are systematically different from those who did not. However, empirical work on the effect of consequentiality scripts in stated preference surveys is in its infancy.

In short, careful study of the incentives at work when people answer stated preference questions helps us understand when such answers *should* be expected to match the behavior that would occur in a real payment situation. In hypothetical surveys, respondents must be faced with an incentive-compatible instrument and must believe the survey to be consequential, both in terms of affecting the provision of the good and in terms of creating a binding payment commitment. If a stated preference study does not satisfy the conditions under which responses should be expected to match those of a real exchange, then an observed mismatch should not be counted as evidence against the efficacy of stated preference methods. Of course, the corollary is also true: if these conditions are met, then a mismatch provides strong evidence of failings in the method.

Lessons From Behavioral Economics: Are the Challenges Unique to Stated Preference?

Most economists use the neoclassical paradigm of rational, optimizing agents to analyze observed outcomes, including survey responses. The last two decades, however, have seen the emergence of behavioral economics—a competing paradigm that seeks to explain persistent departures from neoclassical predictions. This raises a question for stated preference methods: if behavioral anomalies are observed in stated preference outcomes, is it because of a failure of the stated preference method or a failure of the neoclassical paradigm to supply correct predictions for comparison? In this section, we describe findings from research in behavioral economics that need to be considered when we evaluate the accuracy of stated preference methods.

The findings of behavioral economics can be grouped into two broad categories: 1) individual preferences may not be well-behaved in the neoclassical sense and/or 2) individuals do not always optimize when making choices. Departures from neoclassical preferences come in many guises. One example that is particularly relevant for stated preference is the endowment effect, which predicts that people require more compensation to part with something already in possession than they would give up to newly acquire it. This can explain the large divergence in willingness to pay and willingness to accept that is often observed in stated preference surveys, and which is sometimes cited as evidence of the method's failings.

A further example concerns "warm glow," which is the name given to the private value a person receives from the action of contributing to a worthy cause beyond the actual value of the good the contribution provides. The role of warm glow has been hotly debated in the stated preference literature, and its existence was cited by Diamond and Hausman (1994) as a major deficiency in the contingent valuation method. Warm glow is now understood to be one of many reasons for pro-social behaviors such as contributing to public goods (Schokkaert 2006). Social norms and other-regarding preferences such as altruism and reciprocity can also lead individuals to value an environmental good more than its private benefits, in hypothetical as well as real settings. Finally, new results on choices under uncertainty, such as over-weighting small probabilities, are almost certainly relevant for understanding how people respond to survey questions about environmental programs since environmental outcomes are generally uncertain.

Departures from optimizing behavior can also occur for several reasons. We highlight two that are particularly relevant for valuing public goods. First, people may make "mistakes" in general due to bounded rationality and bounded self-control. For example, in the theory of mental accounting (Thaler 1990), money is not fungible across all categories of expenses, meaning multiple budgets constrain different types of behavior. Payments for environmental services in this context are not necessarily constrained by the overall budget, but instead by an expense category that may be more or less binding than fully rational optimization would imply. Li, Berrens, Bohara, Jenkins-Smith, Silva, and Weimer (2005) offer a piece of evidence for mental accounting in contingent valuation: They found that respondents had lower willingness to pay for reduction of global warming when they received reminders about their discretionary income and its use for environmental causes, compared with when they received reminders about their household budget only.

Second, rationality may be the result of repeated participation in markets, where mistakes are costly and individuals learn, rather than an intrinsic characteristic of individual decisionmakers. Departures from rationality can therefore be aggravated by complex or unfamiliar decision environments and uncertainties, which often result in rule-of-thumb behaviors (Iyengar and Kamenica 2007). Although such departures are prevalent in experiments and in field studies of individual choices, stated preference surveys might be more prone to anomalies for two reasons: choices in inconsequential surveys might not be salient and not subject to regulation by institutions, and survey respondents might not have much experience with the environmental goods being valued or with the choice circumstances. However, such anomalies can be alleviated by consumer experience (Whitehead, Bloomquiest, Hoban, and Crawford 1995; List 2003), and perhaps by competitive institutions (Slembeck and Tyran 2004). For example, Cherry, Crocker, and Shogren (2003) showed that market-induced rationality spills over to nonmarket valuations: subjects disciplined by real market-like arbitrage showed lower rates of preference reversals, and the reduced rates carried over to hypothetical settings with money as well as wildlife lotteries.

These developments in behavioral economics offer a richer set of testable hypotheses and interpretations of evidence in contingent valuation studies. The alternative paradigm may be useful for explaining the highly heterogeneous and sometimes nonrational individual outcomes observed in stated preference surveys and experiments, even when aggregate outcomes conform to expectations. In this sense, behavioral insights are useful for providing input into the design and evaluation of stated preference surveys (Shogren and Taylor 2008). However, these new theories also raise fundamental questions about validity tests and research design. For instance, if choices are context dependent, preferences formed in exchange institutions might differ from those formed in nonmarket settings (Bowles 1998). This observation casts doubt on the standard practice of comparing estimates from surveys with those from market data, and it challenges the presumption that the latter should automatically be preferred for use in policy analysis and damage assessment, given that some values are not formed from markets. The conundrum is that one must choose a behavioral paradigm first—for example, behavior based on neoclassical preferences or behavior based on reference-dependent preferences—and then design and implement a study to test the accuracy of a stated preference estimate based on that paradigm. If the findings of the accuracy test are negative, this may provide evidence that the stated preference method is inaccurate *or* that an incorrect behavioral paradigm was chosen.

Empirical Evidence on Validity

How can we assess the empirical accuracy of stated preference methods? In most instances there is no observable "true" value against which an estimate can be judged, and so researchers have devised other means of looking at the accuracy of their estimates. Using definitions from the American Psychological Association, Mitchell and Carson (1989) introduced the concept of "validity" in the context of stated preferences. The validity of a method is essentially the degree to which it correctly measures the theoretical construct under consideration. Table 1 contains a summary of the validity concepts that have now become standard in the literature. A generic definition of each type of validity is provided in question form in the second column, and in the third column we present an example of the question in the specific context of assessing the validity of stated preference studies. We consider each type of validity in turn.

Criterion Validity: Do Stated Preferences Estimates Match Real Payments?

Tests for criterion validity compare the prediction from a stated preference exercise to a standard that is thought to be a suitable proxy for the true measurement objective, which typically involves real payments. In many ways, this validity concept is the most central and salient. Criterion validity has mainly been assessed in the literature using experimental methods in the laboratory and field, but there are also a small number of studies that have timed stated preference studies to coincide with an actual binding referendum.

Table 1
Summary of Validity Concepts for Stated Preference Methods

Criterion	*Generic question*	*Specific question*
Criterion validity	Does the measure relate favorably to other measures that are considered legitimate criteria (i.e., are believed to be accurate)?	Is the estimate generated by stated preference methods the same as a willingness-to-pay value that would be generated if real payment was made?
Convergent validity	Does the measure correlate well with other measures of the same thing?	Is the estimate generated by a stated preference method the same as the willingness-to-pay value that is estimated from a revealed preference method?
Construct validity	Does the measure correlate as expected to other measures as predicted by theory?	Does the estimate generated by a stated preference method relate to income, prices, and other variables in the way economic theory predicts?
Content validity	Does the measure adequately cover the construct's domain?	Does the estimate arise from the best study design practices—including scenario description, econometric analysis, elicitation format, follow up questions, etc.?

Two types of laboratory experiments have been used to gauge criterion validity. In the first, participants are assigned a value for the experimental good as part of the research design. This design allows the researcher to know with certainty the criterion against which real and hypothetical statements of value are compared. Because the value is assigned to the respondent, as opposed to it having arisen internally from the respondent's own preferences, this is known as an "induced value experiment." An advantage of this protocol is that it allows one to focus on value elicitation, as distinct from value formation. Induced value experiments have primarily been used to examine the accuracy of hypothetical referendum-style elicitation vehicles relative to binding real payment votes (for example, Taylor, McKee, Laury, and Cummings 2001; Vossler and McKee 2006; Murphy, Stevens, and Yadav 2010). The results generally show that the distribution of values from hypothetical votes matches the induced-value criterion in aggregate. These findings suggest that a necessary condition for stated preference criterion validity is met. Specifically, when we abstract from the value formation step, there is robust evidence that individuals can be induced to reveal their private willingness to pay for a public good in a properly designed hypothetical situation.

In the second type of experiment, participants' actual values for a real commodity are used as the criterion. These are known as "homegrown value experiments" because participants' own (or homegrown) preferences are the basis for establishing the standard for comparison. In the typical experiment, the criterion is established by a real payment mechanism. For a public good, this takes a referendum format in which all participants must pay a given amount if a majority

votes in favor. The results from hypothetical elicitation formats are then compared to the real payment mechanism as a test of validity. A consistent finding for this type of experiment is that stated values are higher than their real counterparts; this phenomenon has become known as *hypothetical bias*. Meta-analyses by List and Gallet (2001) and Murphy, Allen, Stevens, and Weatherheard (2005) have examined hypothetical bias quantitatively by looking at nearly 30 different lab and field studies that contain both actual and hypothetical estimates of a good's value. List and Gallet find for their sample that hypothetical values exceed actual values on average by a factor of three, while Murphy et al. find the average to be skewed by a few outliers and therefore present a median bias factor of 1.35. More qualitatively, Harrison and Rutström (2008) report that 34 of the 39 studies they surveyed showed upward bias in the hypothetical values. This robust evidence on the existence of hypothetical bias in homegrown value experiments lends support to the notion of criterion *in*validity. The nonvalidity conclusion is also supported by field experiments that include real and hypothetical elicitations for private goods (for example, List 2001; Blumenschein, Blomquist, Johannesson, Horn, and Freeman 2008).

One difficulty in interpreting this set of findings is that not all the studies used in these assessments satisfy the incentive compatibility and consequentiality requirements identified by Carson and Groves (2007) as the necessary conditions for stated responses to match the actual values. For example, Vossler and Evans (2009) find that hypothetical bias disappears from their homegrown value lab experiments when the stated preference elicitation method makes participants feel that their answers are more consequential. Likewise, Landry and List (2007) find that hypothetical bias disappears from their field experiments when respondents are provided with a script emphasizing the consequentiality of the results before answering the value elicitation question. These results jibe well with nonexperimental evidence suggesting that surveys including explicit discussions on how the results might influence policy produce different estimates than those that do not (as in Herriges, Kling, Liu, and Tobias 2010).

Nonetheless, the persistent divergence identified in homegrown value experiments has spawned a large literature dedicated to understanding its causes and finding ways to mitigate its effects. This literature is important for our assessment in that if research can discover a means of eliminating hypothetical bias or predicting its magnitude, the criterion validity of stated preference methods may ultimately be established. For example, one approach is the "cheap talk" method in which participants are explicitly warned of the tendency among people to inflate hypothetically reported values (for example, Cummings and Taylor 1999; List 2001). Over 30 lab and field experiments find that while "cheap talk" can be moderately effective in some circumstances, its net impact varies with the characteristics of participants and the commodity, and the type of script used. The main other alternative, which seems to show more promise, is to calibrate the answers in some way after they have been collected. In one version of this technique, respondents are asked to rate the confidence they have in their answers after completing the elicitation task, which is usually a response to a posted price. Qualitative ranks (for example, "very certain,"

"certain," "uncertain," and do on) as well as multipoint certainty scales have been used, and in most experiments the distribution of hypothetically obtained values can be made to match the distribution of actual values when the uncertain "yes" responses are recoded to "no" responses. Thus, the evidence suggests that one source of hypothetical bias may be in the form of yea-saying by uncertain respondents. Morrison and Brown (2009) provide a summary and reference list of studies related to both the cheap talk, and certainty scale follow-up, methods. Newer vehicles continue to be proposed for minimizing hypothetical bias (Jacquemet, Joule, Luchini, and Shogren forthcoming; Cameron and DeShazo forthcoming; Bateman, Burgess, Hutchison, and Matthews 2008).

A final piece of evidence regarding criterion validity comes from stated preference studies that were conducted in conjunction with actual binding, local referenda. Of these studies, Johnston (2006) is the purest test of criterion validity (and the role of consequentiality) in that the stated preference exercise was executed prior to a local binding referendum and was fielded in an advisory role as input into deciding whether a village in Rhode Island should proceed with the installation of a new water system. Vossler and Kerkvliet (2003) also conduct a survey prior to a binding referendum. Their case study is a 1998 vote over a $9.5 million bond measure, funded by higher property taxes, to pay for improvements to a downtown park in Corvallis, Oregon. In both cases, the researchers find that the stated preference predictions match the outcome of the actual election without any need for calibration. An additional study of this type from Vossler, Kerkvliet, Polasky, and Gainutdinova (2003) found that, if undecided respondents were coded as "no" votes, the stated preference responses were statistically consistent with the referenda results.

How should we interpret the weight of evidence on criterion validity? We have seen that hypothetical bias is commonly found in studies where subjects' personal values form the basis of comparison. On the surface, this provides clear evidence of criterion invalidity for contingent valuation studies. However, a number of steps may be possible to reduce this bias. To the extent that the bias is caused by participants not feeling that their responses matter, stated preference surveys and experiments could be run with designs that provide the proper incentives for subjects to respond thoughtfully. Vossler and Poe (2011) take this a step further when they suggest that criterion validity tests that were conducted without adherence to consequentiality requirements should not be considered when assessing the potential for hypothetical bias. They identify four induced value experiments and one homegrown value experiment that they judge to be consistent with the Carson and Groves (2007) requirements, and note that each of these demonstrates criterion validity. If hypothetical bias remains after appropriate consequentiality conditions are met (or it is not possible to achieve consequentiality), a combination of calibration based on the degree of uncertainty and, to a lesser extent "cheap talk" scripts, might be used to manage hypothetical bias in a way that allows stated preference methods to approach criterion validity status more closely. Finally, the evidence from stated preference surveys and binding referenda supports criterion validity, at least in the

case of people making decisions about local public goods. Based on this string of findings, it is difficult to conclude purely in favor of criterion validity, but also difficult to reject it outright.

For the sake of argument, suppose we find the existing evidence to be insufficient to support a conclusion of criterion validity in the pure sense—that is, statistical equivalence between a stated preference estimate and the criterion. We would still be left with the question as to whether stated preference surveys provide useful (albeit imperfect) information for cost–benefit analysis, policy debates, and/or judicial findings. Indeed, statistical equivalence to one estimate of the truth is a strict standard that many economic analyses used for policy—including most revealed preference estimates, we suspect—would have difficulty passing. More importantly, even limited information may be useful in cost–benefit analysis, policy discussions, and litigation. For example, a simple upper or lower bound on estimates of passive use value can sometimes be sufficient to determine whether a project would pass a cost–benefit analysis. In such a case, a point estimate and knowledge of the direction of bias can be adequate for evaluation. Likewise, even when benefit estimates are uncertain and the sign of any bias is unknown, the magnitude of the point estimate relative to cost estimates (which are also likely to be subject to a range of uncertainties) may provide useful input for policymakers and stakeholders.

Convergent Validity: Are Stated and Revealed Preference Estimates the Same?

Convergent validity refers to how well a stated preference estimate correlates with other measures of the same economic value. The most common type of convergent validity tests compare stated preference estimates to those from other techniques, usually based on revealed preferences. Convergent validity tests of this type are not possible for passive use values, but they can be carried out in other instances, such as when the measurement objective concerns a private or quasi-public good. A good example of this is the value of recreation resources, and many studies have used both stated and revealed preference to examine how the environment conveys value through recreation. If the values match, or diverge in expected directions for expected reasons, the estimates are said to be convergent valid. Of course, both estimates may be wrong! Still, if convergence occurs we might have more confidence in both methods, when they are appropriately applied. In terms of evidence, an older meta-analysis from Carson, Flores, Martin, and Wright (1996) supports the notion of convergent validity. Many individual studies have since been done to study convergent validity between specific types of stated and revealed preference data. In some instances, researchers test for the equivalence of econometric parameters, and in others they test for the statistical equality of economic value estimates. While exceptions exist, our sense is that studies that focus on the equivalence of economic values are generally consistent with the findings from Carson, Flores, Martin, and Wright (1996).

In contemporary research, tests of convergent validity *per se* have given way to a more general focus on econometric methods that allow the two types of data to be combined in the same model to exploit their relative strengths. This literature is

surveyed in a book-length treatment by Whitehead, Haab, and Huang (2011). Here, we merely note that the growth of such methods in environmental and nonenvironmental fields is predicated on the implicit acceptance of convergent validity—or at least a common data-generating process—by a wide spectrum of researchers. Two prominent examples include Berry, Levinsohn, and Pakes (2004), who use both actual purchases and stated intentions to estimate the demand for new car purchases, and Small, Winston, and Yan's (2005) use of both stated and revealed preference data to estimate commuters' demand for travel characteristics. Given this, we interpret the weight of evidence on convergent validity to be generally positive.

Construct Validity: Are Stated Preference Estimates Consistent with Theoretical Predictions?

Prior to the experimental revolution and the advent of research using both stated and revealed preference methods, consideration of construct validity—the extent to which predictions from stated preference experiments are consistent with theory—was the main means by which the efficacy of stated preference was assessed. For example, one issue strongly debated in the 1994 JEP symposium by Diamond and Hausman (1994) and Hanemann (1994) concerns "embedding effects"—that is, whether and to what degree willingness to pay for environmental goods should vary with their size. This has become known as the issue of "scope."

Most of the theory used to evaluate stated preference validity was based on price changes involving private goods, as this was the type of good theretofore most studied by economists. This generated testable predictions and assertions that 1) the proportion of people willing to contribute to an environmental good in a stated preference survey should increase when the requested payment amount falls; 2) people should be willing to pay more to have a higher quantity of the good—that is, estimates should exhibit positive response to scope; 3) the income elasticity of willingness to pay should be larger than one, because environmental quality is best viewed as a luxury good; and 4) willingness to pay and willingness to accept for environmental changes should not be substantially different. While the first of these holds true in almost all stated preference studies, the remaining three were often violated for stated preference data—particularly early studies of sensitivity to scope and most studies comparing estimates of willingness to pay and willingness to accept.

These violations were often cited as evidence of construct invalidity. However, additional work in economic theory since the Exxon spill has shown that predictions 2, 3, and 4 are sensitive to two common features of environmental goods: fixed quantities and limited substitutability with other consumption goods. For example, while the marginal willingness to pay curve for a fixed quantity—like a given level of environmental quality—is downward sloping as expected, its relationship to income imbeds several distinct effects. Flores and Carson (1997) show that the income elasticity of willingness to pay for an environmental good depends on three adjustment margins: the implied income elasticity of demand for the environmental good, the substitutability among all the quantity-constrained goods, and

the share of augmented income allocated to market goods. Numerical examples are used to show that an income elasticity of *willingness to pay* that is less than one is in many plausible circumstances consistent with an income elasticity of *demand* for the fixed quantity that is greater than one. In a similar spirit, Amiran and Hagen (2010) show that bounded substitution between market and environmental goods can result in rational behavior failing to exhibit sensitivity to scope, thereby altering prediction 2 for environmental goods.

Recent empirical results on scope effects deserve mention since the early critiques of stated preference methods were based on findings in some studies that estimates of economic value did not go up when the scale of the environmental good was increased. As sensitivity to scope became a litmus test for the construct validity of stated preference estimates, many post-Exxon studies were specifically designed to include "scope tests." Meta-analyses of these studies from Smith and Osborne (1996), Carson (1997), Brouwer, Langford, Bateman, and Turner (1999), and Ojea and Loureiro (2011) show that scope effects are typically present in well-executed studies.

The persistently observed gap between willingness to pay and willingness to accept estimates in stated preference studies also deserves mention. Although Hanemann (1991) and Zhao and Kling (2009) suggest two different theories that can rationalize such a gap without implying construct invalidity from a neoclassical perspective, the size of the difference in many studies appears implausible. Is the divergence due mainly to the hypothetical nature of stated preference surveys? The evidence suggests no. Horowitz and McConnell (2002) reviewed 45 studies and found no difference in the divergence between hypothetical experiments and real experiments. That is, the divergence is not due to the hypothetical nature of stated preference surveys. Although the divergence has been found to be sensitive to the experimental settings (as in Plott and Zeiler 2005) and experience (as in List 2003), the evidence continues to point to alternative preference structures such as the endowment effect. Thus, the divergence does not automatically translate into violations of construct validity, though it may require reconsideration of what theoretical paradigm is used to analyze behavior.

In sum, advances over the last two decades have shown that a combination of neoclassical and behavioral economic theory can give rise to a wider range of predictions that are consistent with the findings of stated preference studies. Of course, the fact that a wider range of outcomes is theoretically consistent does not validate all possible magnitudes of such outcomes. Even with this caveat, a casual browsing of contemporary state-of-the-art stated preference studies suggests that they are almost always consistent with the predictions noted above. For example, the relationship between the posted price and the probability of a "yes" vote is almost universally negative, income effects are robustly positive, and scope criteria are usually met. The anomalous findings that remain—like the divergence between willingness to pay and accept—arise broadly in other forms of microeconomic data and are therefore of little value in considering the construct validity of stated preference methods.

Nonetheless, as new approaches to stated preference elicitation arise, construct validity concerns can reappear and will need careful attention. For example, the mechanism design framework predicts that ordering effects will be present when individuals respond to multiple choice tasks, as is the case with choice experiments. Ordering effects in choice experiments have indeed been confirmed empirically (Day et al. 2012). Thus, a research challenge is to assess how commonly used departures from incentive compatibility compromise predictions from choice experiments.

Content Validity: Is Best Practice Being Followed?

The final type of validity we consider relates to how effectively a stated preference study adheres to the current state of the art. This topic is relevant for our review insomuch as the notion of state of the art has changed dramatically since the immediate post-Exxon days. The two decades since then have seen an explosion of stated preference work. At a minimum, this means the stock of accumulated wisdom—for example, how people react to a particular payment mechanism, how environmental concepts are best communicated in lay language—is orders of magnitude greater than it was. As mentioned above, there are now several how-to books on stated preference methods that provide survey development steps, numerous examples, and advice on avoiding known pitfalls. Given this, genuine surprises in purely applied studies are now rare; the method has matured and become more standardized, and practitioners now have a much better sense of the important design elements of a stated preference survey.

Evidence for this point is apparent when we look at how the challenges identified in the early debates on the method have been researched and findings incorporated in a new understanding of best practice. We provide three specific examples. First, it is now widely accepted that the environmental good needs to be described with a high level of specificity, and the status quo and changed levels of the good precisely defined in a way that lay respondents can understand and place in context. This information is usually presented via a combination of text, photos, graphics, and numbers that has been deliberately developed using focus groups, interviews, and pretests. The increased use of computer-administered surveys has provided additional flexibility for efficiently explaining the environmental good in multiple ways and checking people's comprehension. A result of this emphasis on specificity (and careful communication) is that contemporary studies almost always satisfy sensitivity to scope and other theoretical predictions. A corollary is that a vague or abstract commodity definition—or inadequate evidence of an effective communication strategy—is considered a failure of content validity. Thus, while the NOAA panel early on stressed the importance of specificity (Arrow, Solow, Portney, Leamer, Radner, and Schuman 1993), its evolution into best practice protocols has occurred incrementally through accumulated experience in numerous subsequent applications.

A second area in which best practice has evolved relates to how the constructed market and payment mechanism are defined and interpreted. It is now widely accepted that the constructed market should represent a realistic mechanism for

bringing about the proposed change, meaning that the size of the change arising from the intervention needs to be seen as physically plausible by respondents. Similarly, the payment mechanism needs to be something that respondents find realistic and familiar—both so they will take the exercise seriously, and so they can envision how an actual payment would occur. The attention given to a survey's policy institutions has also led to a consensus among practitioners that estimated values are for the entire package—that is, the environmental change in the context of the described program, rather than the environmental change in a vacuum. Thus, the expectation among current researchers is not that the estimated values should be independent of context. Instead, differences should arise based on the specifics of the program, and validity hinges on the extent to which the differences are consistent with theory and intuition.

The final example of change in best practice relates to ways that researchers attempt to encourage and/or test for the rationality and truthfulness of respondents' contingent behavior. Understanding of what constitutes an incentive-compatible elicitation mechanism has evolved beyond the NOAA panel's recommendation to use a referendum format (Arrow, Solow, Portney, Leamer, Radner, and Schuman 1993). Researchers now know that design elements related to voluntary versus coercive payment, the actual payment vehicle, and commodity provision details can matter. Likewise, framing the survey to be consequential, the presentation of cheap talk scripts, and the use of certainty follow-up questions have, in various combinations, become common practice. In response to advances in theoretical understanding, researchers are also less likely to draw conclusions about construct validity based on narrowly interpreted tests of scope, income effects, and the sensitivity of value estimates to the details of the constructed market. Instead the criteria used to evaluate construct validity are case-specific and start with questions about the extent to which the specific predictions fit with the specific context.

Content validity is a different concept than the other types of validity in that we cannot summarize general evidence to conclude that stated preference methods are valid or invalid in this dimension. Nonetheless there does seem to be a more complete (and a more nuanced) consensus now than two decades ago on the characteristics of a state-of-the-art study. While this does not say much about the general accuracy of stated preference methods, it does illustrate that the early areas of concern have been well researched and best practice has evolved based on the findings. It is up to the reader to decide if this large volume of work implies we are left with an approach that inspires confidence.

Conclusion

Stated preference techniques are in a much different place in the aftermath of the BP accident in 2010 than they were after the Exxon oil spill in 1989. The past two decades have seen the coming of age of experimental economics, new theoretical developments, accumulating insights from behavioral economics, and a general

Table 2
Summary of Authors' Assessment

Validity concept	*Assessment*	*Comments*
Criterion	Some Yes, Some No	• Persistence of hypothetical bias in homegrown value experiments implies invalidity. • Emerging consequentiality paradigm suggests potential for validity. • Difficult to conclude purely in favor of validity, but also difficult to outright reject validity.
Convergent	Likely Yes	• Formal tests often accept revealed and stated preference equality. Even when statistically different estimates occur, they appear to illustrate common economic phenomena. • Practice has migrated towards using revealed and stated preference data as complements rather than substitutes.
Construct	Strongly Yes	• Further development of standard theory suggests a wider range of outcomes can still be considered neoclassically rational. • New behavioral theories suggest alternative paradigms might be needed to assess validity. • Definitive construct validity tests are now more difficult to formulate.
Content	Variable	• Content validity is a study-specific concept, but the stock of accumulated wisdom suggests adherence to best practice is now a stronger validity concept than in the past.

maturing of the nonmarket valuation literature. We now have more tools with which to judge the accuracy of stated preference estimates and an emerging consensus on the criteria we should use to do so. Many of the questions that arose in the post-Exxon days have been acknowledged and investigated. Those who formulated their beliefs about contingent valuation two decades ago, whether positive or negative, should update their beliefs based on the research agenda that has unfolded. To help readers with this we have prepared Table 2 as our own summary of possible answers to the question of whether the stated preference method can provide valid and accurate estimates of underlying economics values. While the summary constitutes our personal judgments, we have tried to convey the range of views that different people might take following an objective reading of the literature.

Before concluding, we note four areas of research that seem especially critical for continuing the research agenda related to the validity of stated preference methods. First, validity tests that explicitly include the consequentiality dimension in their design are relatively young, and more research is needed to determine if the initial evidence holds up to further scrutiny. Second, much could be learned by subjecting other methods of valuation to the same level of scrutiny that stated preference methods have received. For example, what methods should be used to assess the validity of estimates from hedonic housing or wage studies? How well do recreation demand model estimates stand up to comparisons with actual transactions?

Answers to these questions would enhance their usefulness for cost–benefit analysis generally and improve our ability to assess the relative performance of stated preference methods. Third, a lot of work remains to be done on understanding how the common use of incentive-incompatible designs in choice experiments affects the validity of this recently popular approach. Finally, there remains substantial uncertainty as to how researchers should execute and interpret validity tests using alternative behavioral paradigms. If the same behavioral anomalies appear in both stated and actual behavior, should a valid survey mimic real world choices or seek to elicit "true" preferences—neoclassical or otherwise—for use in welfare analysis?

Despite these and other questions, our sense is that the last 20 years of research have shown that some carefully constructed number based on stated preference analysis is now likely to be more useful than no number in most instances for both cost–benefit analysis and damage assessment. Of course this is a weaker conclusion than validity, and it is not to say that all studies are equally reliable or that inference from reliable studies will always be appropriately applied. But it is illustrative of the remarkable progress that stated preference researchers have made, and it serves as a model for the evaluation of other policy-critical techniques.

■ *The authors appreciate insightful comments from Chang-Tai Hseih, John List, Timothy Taylor, Terry Alexander, Ian Bateman, Trudy Ann Cameron, Richard Carson, Patty Champ, Rick Freeman, Nick Hanley, Joseph Herriges, Jack Knetsch, Rob Johnston, Erin Krupka, Alan Krupnick, John Loomis, Jayson Lusk, Laura Schechter, Jason Shogren, V. Kerry Smith, and John Whitehead. Remaining misinterpretations and errors are the responsibility of the authors alone.*

References

Alvarez, Sergio, Sherry L. Larkin, John C. Whitehead, and Timothy C. Haab. 2012. "Substitution, Damages, and Compensation for Anglers due to Oil Spills: The Case of the Deepwater Horizon." http://purl.umn.edu/124779.

Amiran, Edoh Y., and Daniel A. Hagen. 2010. "The Scope Trials: Variation in Sensitivity to Scope and WTP with Directionally Bounded Utility Functions." *Journal of Environmental Economics and Management* 59(3): 293–301.

Arrow, Kenneth, Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, and Howard Schuman. 1993. "Report of the NOAA Panel on Contingent Valuation." January 11. *Federal Register* 58: 4601–14.

Bateman, Ian J., Diane Burgess, W. George Hutchison, and David I. Matthews. 2008. "Contrasting NOAA Guidelines with Learning Design Contingent Valuation (LDCV): Preference Learning versus Coherent Arbitrariness." *Journal of Environmental Economics and Management* 55(2): 127–41.

Bateman, Ian J., Richard T. Carson, Brett Day, Michael Hanemann, Nick Hanley, Tannis Hett,

Michael Jones Lee, Graham Loomes, Susana Mourato, Ece Ozdemiroglu, David W. Pearce, Robert Sugden, and John Swanson. 2002. *Economic Valuation with Stated Preference Techniques: A Manual*. Cheltenham, UK: Edward Elgar.

Bateman, Ian J., and Kenneth G. Willis, eds. 1995. *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Methods in the US, EC, and Developing Countries*. Oxford, UK: Oxford University Press.

Bennett, Jeff, ed. 2011. *The International Handbook on Non-Market Environmental Valuation*. Cheltenham, UK: Edward Elgar.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112(1): 68–105.

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman. 2008. "Eliciting Willingness to Pay without Bias: Evidence from a Field Experiment." *Economic Journal* 118(1): 114–37.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature* 36(1): 75–111.

Brouwer, Roy, Ian H. Langford, Ian J. Bateman, and R. Kerry Turner. 1999. "A Meta-analysis of Wetland Contingent Valuation Studies." *Regional Environmental Change* 1(1): 47–57.

Cameron, Trudy Ann, and J. R. DeShazo. Forthcoming. "Demand for Health Risk Reductions." *Journal of Environmental Economics and Management*.

Carson, Richard T. 1997. "Contingent Valuation Surveys and Tests of Insensitivity to Scope." Chap. 5 in *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods*, edited by Raymond J. Kopp, Walter Pommerhene, and Norbert Schwartz. Boston: Kluwer.

Carson, Richard T. 2011. *Contingent Valuation: A Comprehensive Bibliography and History*. Cheltenham, UK: Edward Elgar.

Carson, Richard T., Nicholas E. Flores, Kerry M. Martin, and Jennifer L. Wright. 1996. "Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-public Goods." *Land Economics* 72(1): 80–99.

Carson, Richard T., Nicholas E. Flores, and Norman F. Meade. 2001. "Contingent Valuation: Controversies and Evidence." *Environmental and Resource Economics* 19(2): 173–210.

Carson, Richard T., and Theodore Groves. 2007. "Incentive and Informational Properties of Preference Questions." *Environmental and Resource Economics* 37(1): 181–210.

Carson, Richard T., and W. Michael Hanemann. 2005. "Contingent Valuation." Chap. 17 in the *Handbook of Environmental Economics*, Vol. 2, edited by Karl-Goran Maler and Jeffrey Vincent. North-Holland.

Carson, Richard, Robert Mitchell, W. Michael Hanemann, Raymond J. Kopp, Stanley Presser, and Paul Ruud. 2003. "Contingent Valuation and Lost Passive Use: Damages from the *Exxon Valdez* Oil Spill." *Environmental and Resource Economics* 25(3): 257–83.

Champ, Patricia A., Richard C. Bishop, Thomas C. Brown, and Daniel W. McCollum. 1997. "Using Donation Mechanisms to Value Nonuse Benefits from Public Goods." *Journal of Environmental Economics and Management* 33(2): 151–62.

Champ, Patricia A., Kevin J. Boyle, and Thomas C. Brown, eds. 2003. *A Primer on Nonmarket Valuation*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Cherry, Todd L., Thomas D. Crocker, and Jason F. Shogren. 2003. "Rationality Spillovers." *Journal of Environmental Economics and Management* 45(1): 63–84.

Ciriacy-Wantrup, S. V. 1947. "Capital Returns from Soil-Conservation Practices." *Journal of Farm Economics* 29(4, Part 2): 1188–90.

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

Cummings, Ronald G., Glenn W. Harrison, and E. Elisabet Rutström. 1995. "Homegrown Values and Hypothetical Surveys: Is the Dichotomous-Choice Approach Incentive-Compatible?" *American Economic Review* 85(1): 260–66.

Cummings, Ronald G., and Laura O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89(3): 649–65.

Davis, Robert. 1963. *The Value of Outdoor Recreation: An Economic Study of the Maine Woods*. Doctoral dissertation in economics, Harvard University.

Day, Brett, Ian J. Bateman, Richard T. Carson, Diane Dupont, Jordan J. Louviere, Sanae Morimoto, Riccardo Scarpa, and Paul Wang. 2012. "Ordering Effects and Choice Set Awareness in Repeat-Response Stated Preference Studies." *Journal of Environmental Economics and Management* 63(1): 73–91.

de Bekker-Grob, Esther W., Mandy Ryan, and Karen Gerard. 2012. "Discrete Choice Experiments in Health Economics: A Review of the Literature." *Health Economics* 21(2): 145–72.

Diamond, Peter A., and Jerry A. Hausman. 1994. "Contingent Valuation: Is Some Number

Better than No Number?" *Journal of Economic Perspectives* 8(4): 45–64.

***Exxon Valdez* Oil Spill Trustee Council.** N.d. "Oil Spill Facts." Webpage. http://www.evostc.state.ak.us/facts/index.cfm.

***Exxon Valdez* Oil Spill Trustee Council.** N.d. "Settlement." Webpage about the U.S. District Court Settlement on October 9, 1991 between the State of Alaska, the U.S. government, and Exxon. http://www.evostc.state.ak.us/facts/settlement.cfm.

Flores, Nicholas E., and Richard T. Carson. 1997. "The Relationship between the Income Elasticities of Demand and Willingness to Pay." *Journal of Environmental Economics and Management* 33(2): 287–95.

Guarino, Mark. 2012. "BP Oil Spill Settlement: Justice for 100,000 Gulf Coast Victims?" *Christian Science Monitor*, April 18. http://www.csmonitor.com/USA/Justice/2012/0418/BP-oil-spill-settlement-Justice-for-100-000-Gulf-Coast-victims.

Haab, Timothy C., and Kenneth E. McConnell. 2002. *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*. Cheltenham, UK: Edward Elgar.

Hanemann, W. Michael. 1991. "Willingness to Pay and Willingness to Accept: How Much Can They Differ?" *American Economic Review* 81(3): 635–47.

Hanemann, W. Michael. 1994. "Valuing the Environment through Contingent Valuation." *Journal of Economic Perspectives* 8(4): 19–43.

Harrison, Glenn W., and E. Elisabet Rutström. 2008. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." Chapter 81 in *Handbook of Experimental Economics Results*, Vol 1, Part 5, edited by Charles R. Plott and Vernon L. Smith. New York: North-Holland.

Hausman, Jerry A., Gregory K. Leonard, and Daniel L. McFadden. 1995. "A Utility-Consistent, Combined Discrete Choice and Count Data Model Assessing Recreational Use Losses due to Natural Resource Damage." *Journal of Public Economics* 56(1): 1–30.

Herriges, Joseph A., Catherine L. Kling, Chih-Chen Liu, and Justin Tobias. 2010."What Are the Consequences of Consequentiality?" *Journal of Environmental Economics and Management* 59(1): 67–81.

Horowitz, John K., and Kenneth E. McConnell. 2002. "A Review of WTA/WTP Studies." *Journal of Environmental Economics and Management* 44(3): 426–47.

Iyengar, Sheena S., and Kamenica, Emir. 2007. "Choice Overload and Simplicity Seeking." http://www.cbdr.cmu.edu/seminar/Emir2.pdf.

Jacquemet, Nicolas, Robert-Vincent Joule, Stephane Luchini, and Jason F. Shogren. Forthcoming. "Preference Elicitation under Oath." *Journal of Environmental Economics and Management*. Available online at http://www.sciencedirect.com/science/article/pii/S0095069612000587.

Johnston, Robert J. 2006. "Is Hypothetical Bias Universal? Validating Contingent Valuation Responses Using a Binding Public Referendum." *Journal of Environmental Economics and Management* 52(1): 469–81.

Kanninen, Barbara J., ed. 2007. *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Approach to Theory and Practice*. Springer.

Krutilla, John. 1967. "Conservation Reconsidered." *American Economic Review* 57(4): 777–86.

Landry, Craig F., and John A. List. 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): 420–29.

Larkin, Sherry. 2012. Personal communication with the authors.

Li, Hui, Robert P. Berrens, Alok K. Bohara, Hank C. Jenkins-Smith, Carol L. Silva, and David L. Weimer. 2005. "Testing for Budget Constraint Effects in a National Advisory Referendum Survey on the Kyoto Protocol." *Journal of Agricultural and Resource Economics* 30(2): 350–66.

List, John 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, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118(1): 41–71.

List, John A., and Craig A. Gallet. 2001. "What Experimental Protocol Influence Disparities between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics* 20(3): 241–54.

Loomis, John, Paula Kent, Liz Strange, Kurt Fausch, and Alan Covich. 2000. "Measuring the Total Economic Value of Restoring Ecosystem Services in an Impaired River Basin: Results from a Contingent Valuation Survey." *Ecological Economics* 33(1): 103–117.

Louiviere, Jordan J., David A. Hensher, and Joffre D. Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press.

Mitchell, Robert Cameron, and Richard T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Methods*. Washington, DC: Resources for the Future.

Morrison, Mark, and Thomas C. Brown. 2009. "Testing the Effectiveness of Certainty Scales,

Cheap Talk, and Dissonance-Minimization in Reducing Hypothetical Bias in Contingent Valuation Studies." *Environmental and Resource Economics* 44(3): 307–326.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherheard. 2005. "A Meta-analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics* 30(3): 313–25.

Murphy, James J., Thomas H. Stevens, and Lava Yadav. 2010 "A Comparison of Induced Value and Home-grown Value Experiments to Test for Hypothetical Bias in Contingent Valuation." *Environmental and Resource Economics* 47(1): 111–23.

Ojea, Elena, and Maria L. Loureiro. 2011. "Identifying the Scope Effect on a Meta-analysis of Biodiversity Valuation Studies." *Resource and Energy Economics* 33(3): 706–24.

Phaneuf, Daniel J., and Till Requate. Forthcoming. *A Course in Environmental Economics: Theory, Policy, and Practice*. Cambridge University Press.

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 Valuations." *American Economic Review* 95(3): 530–45.

Portney, Paul R. 1994. "The Contingent Valuation Debate: Why Economists Should Care." *Journal of Economic Perspectives* 8(4): 3–17.

Randall, Alan, Berry Ives, and Clyde Eastman. 1974. "Bidding Games for Valuation of Aesthetic Environmental Improvements." *Journal of Environmental Economics and Management* 1(2): 132–49.

Shogren, Jason F., and Laura O. Taylor. 2008. "On Behavioral-Environmental Economics." *Review of Environmental Economics and Policy* 2(1): 26–44.

Slembeck, Tilman, and Jean-Robert Tyran. 2004. "Do Institutions Promote Rationality?: An Experimental Study of the Three-Door Anomaly." *Journal of Economic Behavior and Organization* 54(3): 337–50.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73(4):1367–82.

Smith, V. Kerry, and Laura Osborne. 1996. "Do Contingent Valuation Estimates Pass a Scope Test? A Meta-analysis." *Journal of Environmental Economics and Management* 31(3): 287–301.

Schokkaert, Erik. 2006. "The Empirical Analysis of Transfer Motives." Chap. 2 in *Handbook of the Economics of Giving, Altruism and Reciprocity*, Vol. 1, edited by Serge-Christophe Kolm and Jean Mercier Ythier. North-Holland.

Taylor, Laura O., Michael McKee, Susan K. Laury, and Ronald G. Cummings. 2001. "Induced-Value Tests of the Referendum Voting Mechanism." *Economic Letters* 71(1): 61–65.

Thaler, Richard H. 1990. "Anomalies: Saving, Fungibility, and Mental Accounts." *Journal of Economic Perspectives* 4(1): 193–205.

Vossler, Christian A., and Mary F. Evans. 2009. "Bridging the Gap between the Field and the Lab: Environmental Goods, Policy Maker Input, and Consequentiality." *Journal of Environmental Economics and Management* 58(3): 338–45.

Vossler, Christian A., and Joe Kerkvliet. 2003. "A Criterion Validity Test of the Contingent Valuation Method: Comparing Hypothetical and Actual Voting Behavior for a Public Referendum." *Journal of Environmental Economics and Management* 45(3): 631–49.

Vossler, Christian A., Joe Kerkvliet, Stephen Polasky, and Olesy Gainutdinova. 2003. "Externally Validating Contingent Valuation: An Open-Space Survey and Referendum in Corvallis, Oregon." *Journal of Economic Behavior and Organization* 51(2): 261–77.

Vossler, Christian A., and Michael McKee. 2006. "Induced-Value Tests of Contingent Valuation Elicitation Methods." *Environmental and Resource Economics* 35(2): 137–68.

Vossler, Christian A., and Gregory L. Poe. 2011. "Consequentiality and Contingent Values: An Emerging Paradigm." Chapter 7 in *The International Handbook on Non-Market Environmental Valuation*, edited by Jeff Bennett. Northhampton, MA: Edward Elgar.

Whitehead, John C., Glenn C. Blomquist, Thomas J. Hoban, and William B. Crawford. 1995. "Assessing the Validity and Reliability of Contingent Values: A Comparison of On-site Users, Off-site Users, and Non-users." *Journal of Environmental Economics and Management* 29(2): 238–51.

Whitehead, John, Tim Haab, and Ju-Chin Huang, eds. 2011. *Preference Data for Environmental Valuation: Combining Revealed and Stated Approaches*. Routledge Explorations in Environmental Economics series. Routledge.

Zhao, Jinhua, and Catherine L. Kling. 2009. "Welfare Measures when Agents Can Learn: A Unifying Theory." *Economic Journal* 119(540): 1560–85.