CONTEMPORARY GUIDANCE FOR STATED PREFERENCE STUDIES

Abstract: This article proposes contemporary best-practice recommendations for stated preference (SP) studies used to inform decision-making, grounded in the accumulated body of peer-reviewed literature. These recommendations consider the use of SP methods to estimate both use and non-use (passive-use) values, and cover the broad SP domain including contingent valuation and discrete choice experiments. We focus on applications to public goods in the context of the environment and human health, but also consider ways in which the proposed recommendations might apply to other common areas of application. The recommendations recognize that SP results may be used and reused (benefit transfers) by governmental agencies, non-governmental organizations, and that all such applications must be considered. The intended result is a set of guidelines for SP studies that is more comprehensive than that of the original National Oceanic and Atmospheric Administration (NOAA) Blue Ribbon Panel on contingent valuation, is more germane to contemporary applications, and reflects the two decades of research since that time. We also distinguish between practices for which accumulated research is sufficient to support recommendations and those for which greater uncertainty remains. The goal of this article is to raise the quality of SP studies used to support decision making and promote research that will further enhance the practice of these studies worldwide.

JEL Codes: C83, H41, Q51

Keywords: choice experiment, choice modelling, contingent valuation, guidelines, non-market valuation, questionnaire, stated preference, survey

1. INTRODUCTION

Stated preference (SP) methods estimate measures of economic value using responses to survey questions. Multiple variants of these methods exist. One common approach is discrete choice contingent valuation (CV), wherein respondents are asked whether they would vote for a proposed change at a specified cost. Another common example is the discrete choice experiment (DCE or simply CE), in which respondents are asked to indicate their preference among two or more multi-attribute alternatives.[1] SP methods such as these represent the only known approach to estimate values for changes in many public goods including environmental services, human health effects, and other outcomes for which (direct or indirect) revealed preference (RP) data are not available. Hence, they are the only available means to estimate non-use (also called nonuse or passive-use) values, or use values associated with changes that fall outside the range of current markets or observed conditions. SP methods thus have a unique role in welfare analysis.

SP methods are also controversial, with debates over their validity magnified by highly publicized cases such as damage assessments for the 1989 *Exxon Valdez* oil spill in the US (Carson et al. 2003) and proposed mining activities adjacent to Kakadu National Park in Australia (Bennett 1996). Three articles in the *Journal of Economic Perspectives*, two decades removed from these early applications, provide an overview of the continuing controversy (Carson 2012; Hausman 2012; Kling, Phaneuf, and Zhao 2012).[2] These and other articles reflect

[1] In this article, we use CV to distinguish studies that estimate values for a change or set of changes considered as an indivisible whole, and CEs to identify studies that define the change(s) to be valued as a function of multiple attributes (or characteristics) that may each take different levels. As defined here, the CV versus CE distinction is hence akin to a non-attribute versus attribute-based one. We recognize that not all practitioners agree that this is the most useful or precise terminology (for an alternative see Carson and Louviere 2011), but our use of this nomenclature follows the most common terminology in the literature. Many of the findings and lessons from the CV literature are relevant to the CE literature and vice versa, and so we make a distinction only where relevant.

[2] While this debate has focused on SP methods, these methods are characterized by several anomalies that also occur in markets, where market data are given *prima facie* credibility (e.g., Hanemann 1994; Bateman et al. 1997a,b, 2008b; Carson 2012).

the intense debate over whether SP methods can provide credible information to inform decision-making. Particular attention has been given to the issue of hypothetical bias, or whether values estimated using SP data are equivalent to those that would be estimated using parallel RP data (in cases where valid comparisons are possible; we discuss this issue below). Despite this controversy, the need for information on economic values in the absence of market choices leads to an unquestionable demand for SP value estimates.

SP techniques were first used widely during the 1980s and 1990s. Starting from a few dozen published papers and two key books (Cummings, Brookshire, and Schulze 1986; Mitchell and Carson 1989), the literature expanded to at least 7,500 SP studies in the published and grey literatures by 2011 (Carson 2011; Kling et al. 2012), with more since that time. Results of SP studies are central components of formal and informal policy analyses, natural resource damage assessments and other types of litigation, decision-making by firms, and advocacy by NGOs. This breadth of applications reflects a continuing expansion in the use and relevance of SP results. SP methods have become ensconced in the valuation of changes in environmental goods such as ecosystem services, as well as in transportation, health, marketing and other applications.

Over two decades ago, the NOAA Blue Ribbon Panel report on CV (Arrow et al. 1993) stimulated a research agenda that fundamentally influenced the design and conduct of SP studies, particularly within the context of environmental valuation. The Panel focused on the use of CV to estimate non-use values for litigation in the United States, and proposed what they referred to as "a fairly complete set of guidelines compliance with which would define an ideal CV survey." These guidelines spurred research to advance the validity and reliability of CV methods and were an indirect impetus for the expanding use of CEs.

SP research has advanced considerably since the NOAA Panel, and this evolving

research affects the applicability of the NOAA Panel's guidelines. Some of the Panel's guidelines have proven to be germane and fundamental (e.g., the framing of the environmental value elicitation as an advisory public referendum). Other concerns, such as the ability of the method to demonstrate scope effects, have been largely established in the literature. Still other guidelines are subject to question. An example is the recommendation to use personal interviews, a guideline that was never widely adopted and did not consider the subsequent emergence of internet surveys. Furthermore, CEs, a now-common technique, were not considered. Despite this evolution of research and practice, many evaluations continue to use the NOAA Panel guidelines as the primary criteria upon which to evaluate SP studies. A consequence of reliance on arguably outdated guidance is adherence to norms for best practice that are no longer (or are questionably) relevant, which can lead to inconsistencies in the quality and practice of SP studies used to inform decisions. Relying on dated guidance can also have unintended influences on the state and direction of methodological research.

More recent guidance, including that in the academic literature and issued by government agencies in the US, the EU and Australia, has neither developed nor provided contemporary recommendations concerning best practices (e.g., Carson 2000; Pearce et al. 2002; Swedish EPA 2006; Riera et al. 2012; Baker and Ruting 2014; Freeman, Herriges and Kling 2014; US EPA 2014). For the most part, these documents describe basic methods and are often equivocal (e.g., describing issues and techniques rather than specifying best practices), deferring to the NOAA Panel, or simply listing issues to consider as part of quality evaluations.

Of particular concern are perceived conflicts between recommendations in guidance documents and evidence from the contemporary peer-reviewed literature, and the lack of guidance on important design and implementation features. For example, the evolving weight of

evidence in the literature suggests that incentive compatibility is important for validity (Carson and Groves 2007; Collins and Vossler 2009; Taylor, Morrison, and Boyle 2010; Vossler, Doyon, and Rondeau 2012; Carson, Groves, and List 2014).[3] Yet existing guidance documents frequently allow or endorse value elicitation mechanisms such as classic open-ended questions that are known *not* to be incentive compatible, or imply (incorrectly) that some types of questions are incentive compatible regardless of other aspects of the elicitation (e.g., type of good and payment vehicle; perceived consequentiality).[4] Examples such as these suggest the need for contemporary guidance that better reflects recent and accumulated findings from the SP literature.

This article proposes contemporary best-practice recommendations for SP studies intended to inform decision-making, grounded in the accumulated body of knowledge from the peer-reviewed literature. These recommendations consider (a) the use of SP methods to estimate both use and non-use values, (a) the comprehensive SP domain including both CV and CEs, and (c) a broad spectrum of contemporary applications. We focus on applications to public goods in environmental and human health contexts, but consider ways in which the proposed recommendations might apply to other common areas of application (e.g., private goods in transportation and marketing). The proposed recommendations recognize that SP results are often used and reused (benefit transfers) by government agencies, NGOs and others, and that all such uses must be considered. We also distinguish between practices for which accumulated

[3] An incentive compatible mechanism is one in which the respondent theoretically has the incentive to truthfully reveal any private information asked for by the mechanism or that truthful preference revelation is the dominant strategy (Carson et al. 2014).

[4] Consequentiality describes a condition in which an individual faces or perceives a non-zero probability that their responses will influence decisions related to the outcome in question and they will be required to pay for that outcome if it is implemented. Consequentiality is one component of incentive compatible value elicitation, and as such is a necessary but not sufficient condition (Carson et al. 2014; Herriges et al. 2010; Vossler et al. 2012).

research is sufficient to support recommendations and those for which greater uncertainty remains (and hence additional research is required to support recommendations). The recommendations focus on major decisions that must be made as part of any SP study, rather than specific implementation steps or technical minutiae. Finally, we take a global perspective that looks beyond the narrow domain of non-use values for US litigation, which was the focus of the NOAA Panel report, towards broader worldwide SP applications that are now common.[5] The overall goal of this article is to raise the quality of SP studies used to support decision-making and promote research that will further enhance the practice of these studies worldwide.[6]

2. BRIEF HISTORY OF SP METHODS

It is often challenging to identify who was the first to propose or implement an empirical method, but there is general consensus regarding the set of influential contributions that spawned the SP literature. The basic conceptual framework for SP methods can be traced back to Thurstone (1927). Two decades later, Ciriacy-Wantrup (1947) proposed asking people to state values for items that are not traded in markets. The first known empirical application of CV was by Davis (1963). CEs hail back to their precursor, conjoint analysis, and the work of Luce and Tukey (1964). Empirical applications evolved in marketing (Green and Rao 1971; Green and Srinivasan 1978). McFadden (1986) demonstrated how choice-based conjoint elicitations in marketing could be analyzed using econometric tools for discrete choice analysis. One of the first cross-overs to environmental economics is found in research by Lareau and Rae (1989). Both CV and CEs have theoretical foundations in Lancaster's (1966) approach to consumer

[5] We do not identify the extent to which individual recommendations are the same, differ or are new relative to the NOAA Panel guidelines. Any point-by-point comparison is left to the reader.
[6] As an additional benefit, the increased methodological consistency engendered through guidelines may enable more informed comparisons across studies.

theory, in which goods and services are comprised of bundles of attributes that consumers value. Many SP methods have empirical common ground in random utility modeling (Manski 1977).

Several milestone events influenced the evolution of the SP literature. Early CV applications began to emerge in the US in the late 1960s and Europe by the early 1970s (Mitchell and Carson 1989; Bonnieux and Rainelli 1999). Australian applications emerged later, with 12 studies identified as of 1979 (Bennett 1996). The same year, Bishop and Heberlein (1979) undertook one of the most widely cited early criterion validity investigations of CV, notable for its comparison of SP estimates to those generated via an experiment with actual cash payments.[7] At that time, many economists were of the early mindset of Scott (1965) who commented: "ask a hypothetical question and you will get a hypothetical response" (p. 37). Bishop and Heberlein (1979) addressed this assertion with empirical evidence that is often interpreted as giving credibility to SP estimates.[8] A few years later Cummings et al. (1986) edited a book on CV that integrated the combined wisdom of leading academics on the state of the art. This assessment led to four reference operating conditions (p. 107) to enhance CV applications (Kahneman added three additional operating conditions in the same book, p. 186-94). This work provided the first published guidance for CV. In 1989 the US Court of Appeals ruled that compensable values include "option and existence values"[9] which opened the door for CV estimates to be used in litigation, since only SP methods can measure non-use values. The same year, Mitchell and Carson (1989) published *Using Surveys to Value Public Goods: The CV Method*, which was the

[7] Davis (1963) and others had previously compared SP results to those generated via RP methods such as travel cost and property value analyses as part of convergent validity investigations (Mitchell and Carson 1989).
[8] This study compared SP responses and responses to actual cash offers for goose hunting permits (a private good) to estimates from a travel cost (TC) recreation demand model. Among other findings, results suggested that CV WTP estimates were similar to TC estimates, both of which understated the actual cash amount at which hunters were willing to sell permits. SP willingness to sell overstated actual willingness to sell by approximately 60%.
[9] 880 F.2d 432, 279 US App. D.C. 109, p. 44.

first detailed guide for the design and implementation of this SP method.

The evolution of SP methods was altered permanently with the *Exxon Valdez* oil spill. After the natural resource damage claim settlement was approved in 1991 (http://www.evostc.state.ak.us/index.cfm?FA=facts.settlement, accessed September 22, 2016), consultants employed by Exxon compiled evidence suggesting that CV could not reliably measure non-use values (Hausman et al. 1993). The NOAA Administrator, the lead federal trustee for the Exxon Valdez Natural Resource Damage Assessment, countered with an expert panel that led to the NOAA Panel Report on Contingent Valuation (Arrow et al. 1993). These events promoted extensive research to investigate SP validity.[10] They also stimulated the emergence of CEs in environmental and health applications (Boxall et al. 1996; Adamowicz, Louviere, and Williams 1994, 1998; Hanley et al. 1998; Ryan 1999). Since that time, there have been multiple books published on SP methods (e.g., Bateman and Willis 1999; Louviere, Hensher, and Swait 2000; Bennett and Blamey 2001; Bateman et al. 2002; Champ, Boyle, and Brown 2003; Ryan, Gerard, and Amaya-Amaya 2008; Kanninen 2010), along with articles summarizing SP methods in different areas of economics (e.g., Carson 2000; Boxall et al. 1996; Hanley, Mourato, and Wright 2001; Ryan and Gerard 2003; Carson and Hanemann 2005; Hoyos 2010; de Bekker-Grob et al. 2012). In the intervening years, however, there has been neither a synthesis of best-practice recommendations akin to those enumerated in Cummings et al. (1986) and Arrow et al. (1993), nor a synthesis that integrated best practices for both CV and CEs.

3. FOUNDATIONS FOR THE RECOMMENDATIONS

The recommendations presented here have been developed based on a review of the literature on

[10] List and Gallet (2001) discuss and meta-analyze many of these pre-2000 studies.

SP methods applied in multiple fields, including environmental economics, health economics, transportation economics and marketing. This review has been combined with a formal, multi-year process through which input was solicited from SP practitioners, consumers of SP research, and other interested parties. In response to these solicitations, formal written input was provided by over two dozen individuals and groups. In-person input was also received during and after presentations of draft guidelines at professional meetings.[11] Finally, we have drawn from the combined experience of the authors regarding the design and implementation of SP studies

The presented recommendations are grouped into five categories: (a) survey development and implementation, (b) value elicitation, (c) data analysis, (d) validity assessment, and (e) study reporting. As emphasized above, the goal of this paper is to propose a set of best practices for the design and implementation of SP studies used to *support decision-making*. Among the motivations is to reduce uncertainty surrounding the use of SP methods to inform decisions, and to assist researchers, practitioners, reviewers, users and funders to understand best practices when considering designing, implementing or using the estimates from a SP study. Consensus recommendations supported by the best available research can help develop understanding and acceptance of any empirical method by reducing unnecessary heterogeneity in application practice. *These recommendations are NOT meant to impose prescriptive and mandatory constraints on SP research or publications, as not all recommendations may apply in all*

[11] Elements of the process included: (1) an interactive session at the 2014 World Congress of Environmental and Resource Economists in Istanbul, Turkey, (2) formal solicitation of written input and feedback distributed through multiple professional organizations worldwide, including the Association of Environmental and Resource Economists (AERE), the European the Association of Environmental and Resource Economists (EAERE), the Australian Agricultural and Resource Economics Society (AARES), the Land and Resource Economics Network (ResEcon), the International Academy for Health Preference Research (IAHPR), the UK Health Economists' Study Group (HESG), and others, (3) discussions with government agency staff and scientists at formal and informal meetings, (4) presentation and discussion of draft guidance at organized sessions at annual meetings of the International Choice Modeling Conference (ICMC), AERE, and EAERE (all in 2015), and (5) follow-up discussions with experts in particular areas of SP methodology.

research contexts. In fact, we expect this paper to spur needed research in areas where the recommendations may be subject to debate, or where they remain incomplete.

4. SURVEY DEVELOPMENT AND IMPLEMENTATION

SP questionnaire design should follow best practices applicable to all types of survey research (see Bateman et al. 2002; Lancsar and Louviere 2008; Groves et al. 2009; Kanninen 2010; Bridges et al. 2011; Rossi et al. 2013; Dillman, Smyth, and Christian 2014; Champ, Boyle, and Brown 2017). Among the goals of SP survey development is to maximize the validity and reliability of the resulting value estimates. As commonly used in the literature, validity refers to the minimization of bias in estimates, while reliability refers to the minimization of variability (Bateman et al. 2002; Mitchell and Carson 1989; Bishop and Boyle 2017). Good survey design and implementation procedures are crucial to accomplishing these goals and are necessary if we wish to extrapolate model estimates, based on a survey sample, to an intended population.

Choices of effective methods for survey design, sampling and data collection can enhance the validity and reliability of SP value estimates. At a basic level these choices include: (a) designing a survey instrument that clearly explains baseline (or status quo) conditions and poses a consequential valuation question; (b) selecting a random sample of the potentially affected population; and (c) choosing a survey mode with desired properties. This section recommends design and implementation procedures that are consistent with general survey research practice, with emphasis on elements that are unique to SP surveys.

4.1. Scenario Descriptions

SP methods provide estimates of value associated with changes in economic welfare brought

about by a change in the world, measured from a particular baseline. The baseline (or status quo) condition(s), as well as the proposed change(s) relative to the baseline, must be described in a way that is understood and viewed as credible by respondents, and that enables respondents to anticipate accurately the likely effects on their welfare. Meeting these conditions requires survey design procedures to ensure subjects' understanding and acceptance of valuation scenarios.

RECOMMENDATION 1: SP questionnaires should clearly present the baseline (or status quo) condition(s), the mechanism of change, and the change(s) to be valued, and should elicit evidence that these pieces of information are understood, accepted and viewed as credible by respondents. Both objective information and subjective (respondent) perceptions of this information should be considered. Temporal, spatial, uncertainty and risk dimensions, and whether the baseline and change(s) are individual- or household-specific, should be identified.

Questionnaires must describe the type and extent of change for each valuation scenario, using language that is accurate and understandable by respondents. Scenario descriptions also require information on the mechanism or set of mechanisms (usually a policy revision, investment, or management change) that will generate the change(s) to be valued. Mechanisms should be described, to the extent practical, in a way that is consistent with a plausible real-world action that would bring about the change.[12]

Baselines and changes should be presented in accurate, measureable and interpretable terms (Schultz et al. 2012) and should reflect outcomes for which utility consequences can be clearly identified by respondents (Boyd and Krupnick 2013; Johnston et al. 2016). Imprecise or qualitative terms such as "high," "medium" and "low" should be avoided unless these terms are

[12] Scenario design should also recognize that elicited WTP can be influenced by the mechanisms or processes used to achieve changes (Bulte et al. 2005; Johnston and Duke 2007; Bosworth, Cameron, and DeShazo 2010).

clearly defined and understood by respondents (Johnston et al. 2012). If intermediate changes[13] (or processes to accomplish changes) are to be valued, respondents should be able to identify the linkages between these intermediate changes and the "final" changes that are directly relevant to welfare, enabling them to understand the consequences that affect their welfare (Boyd and Krupnick 2009, 2013; Boyd et al. 2016; Johnston et al. 2016).

Baseline conditions and changes, as well as other components of the valuation scenario, may include subjective perceptions by respondents (e.g., perceived versus actual health risks) that require elicitation and communication to avoid scenario rejection or adjustment (Adamowicz et al. 1997, 2014; Cameron, DeShazo, and Johnson 2011). Scenario design must take subjective perceptions into consideration when describing the actual change to be valued.[14] This includes presenting information in a manner that subjects will understand and accept, and using deliberate strategies to evaluate whether respondents are adjusting or reinterpreting the presented (actual) information according to their subjective perceptions of the scenario.

Scenario design must therefore consider the impacts of (a) the provided information on subjects' responses to valuation questions (Bergstrom, Stoll, and Randall 1989; Boyle 1989; Ajzen, Brown, and Rosenthal 1996; Blomquist and Whitehead 1998; Hoehn and Randall 2002), (b) the framing of valuation questions (Rolfe, Bennett, and Louviere 2002), (c) sequencing, if more than one valuation question is included in a questionnaire (Alevy, List, and Adamowicz 2011; Day and Prades 2010; Day et al. 2012), and (d) respondents' prior experience and knowledge (Cameron and Englin 1997; LaRiviere et al. 2014). Failure to consider and address these issues can lead to invalid or unreliable welfare estimates. Both the quantity and type of

[13] An example of an intermediate change would be a change in wetland processes valued primarily due to impacts on other outcomes, such as increases in wetland-dependent species or flood attenuation.
[14] For example, one should not assume, without pretesting, that lay respondents accept and understand all objective information in a questionnaire in the same way as experts accept and understand the same information.

information provided are relevant. Information required to describe the baseline and change may involve spatial and temporal features that require extra effort to explain[15] so that subjects understand the valuation scenario and its relevance (Johnston, Swallow, and Bauer 2002, 2016; Bateman et al. 2005, 2006a; Horne, Boxall, and Adamowicz 2005; Boyle et al. 2010; Meyer 2013).[16]

Among the most fundamental design features of SP scenarios is the bid (or cost) amount posed for the object of choice. These monetary amounts must be clearly stated[17] along with who pays (e.g., household or individual), whether payments are mandatory or voluntary,[18] the frequency of payment (e.g., annual or monthly), the duration of payment (e.g., one time or annually for five years), and the method of payment (payment vehicle, e.g., income tax or utility bill). Amounts and payment vehicles must be credible and salient to respondents, and must cover (i.e., be potentially paid by) a sufficient proportion of the sampled population to enable extrapolation of results to that population. Finally, the features of the payments must be consistent with mechanisms described to bring about the change to be valued.

In many cases, baselines or changes are not known with certainty. Here, we refer to uncertainty in the baselines or changes that will occur *within each valuation scenario*, beyond that which can be adequately captured via approaches such as the use of multiple attribute levels in a CE or the treatment of outcomes as expected values. Within SP studies, this is often conceptualized as a case of risk, in which a probability distribution of possible states-of-the-

[15] For example, questionnaires might include maps of changes that are projected under alternative future scenarios.
[16] Note that the sensitivity of a welfare estimate to differences in factors such as information provision or framing does not necessarily imply lack of validity or reliability. As in RP contexts, behavior in SP contexts can vary (according to factors such as these) in ways that are consistent with valid and reliable welfare estimation.
[17] If an elicitation mechanism does not provide a predetermined amount (e.g., open-ended elicitation), all other aspects of the payment must be clearly described.
[18] As discussed below, mandatory (or binding) payments are required for incentive compatibility.

world is known or can be approximated. A growing literature illustrates the relevance of risk and uncertainty (e.g., over the change in an environmental outcome) for SP studies (e.g., Roberts, Boyer, and Lusk 2008; Hanley, Kriström, and Shogren 2009; Shaw and Baker 2010; Cameron et al. 2011; Glenk and Colombo 2011, 2013; Akter, Bennett, and Ward 2012; Rolfe and Windle 2015), following similar relevance for welfare analysis in general (Graham 1981). This work demonstrates that the omission of relevant risk information from scenarios (i.e., treating risky outcomes as certain) can lead to empirical value estimates that do not reflect *ex ante* welfare change under uncertainty. Hence, when risk or uncertainty is an important aspect of the baseline or change being valued, scenarios should communicate this information in terms that are readily understood by respondents (Lundhede et al. 2015).[19] As above, survey designers should consider the possibility that subjects will use subjective perceptions to modify objective risk information in scenarios (Cameron 2005b, 2011; Lee and Cameron 2008; Cai, Cameron, and Gerdes 2010; Adamowicz et al. 2014).

The sufficiency of these and other design components depends on respondent understanding and perceptions. Hence, all should be tested using processes both external to the questionnaire (e.g., focus groups, one-on-one tests, verbal protocols) and internal (e.g., debriefing questions assessing understanding), as described below.

4.2. Survey Pretesting

The survey literature documents procedures for survey pretesting (Presser et al. 2004), but

[19] Although the literature has not settled on best practices to communicate risks in SP questionnaires, past work provides insights into effective and potentially ineffective methods (Fischhoff, Brewer, and Downs 2011; Harrison et al. 2014). Approaches shown to be effective include risk tutorials and visual depiction of risks using ladders or grids, often accompanied by verbal or numeric statements (Lipkus and Hollands 1999; Corso, Hammitt, and Graham 2001; Burr et al. 2012; Adamowicz et al. 2014; Viscusi 2014). Some SP studies have compared different risk communication approaches (e.g., Loomis and duVair 1993; Corso et al. 2001; Botzen and van den Bergh 2011).

specific and comprehensive guidance as to what pretest procedures should be used for SP questionnaires is rare. This lack of guidance aside, the quality of a survey instrument relies on pretesting, and pretesting is a central component of content validity (Smith 2006b; Carson 2012).

RECOMMENDATION 2: Qualitative pretesting is a necessary component of survey design. Whether focus groups and/or cognitive interviews are employed, and the appropriate number of each, varies across contexts. For most applications, a rough minimum of four to six focus groups is recommended, with larger numbers recommended for new, unfamiliar or difficult-to-quantify goods. Quantitative pretesting using data from pilot studies should be conducted, where feasible, to facilitate bid and attribute designs for SP questions, calibrate survey and item response rates, and conduct preliminary statistical tests of hypotheses. All types of pretesting should be conducted using members of the target population whenever possible. Survey designers should archive their pretest records, including scripts used to administer focus groups or interviews, number of focus groups or interviews, characteristics of participants, subject-selection methods, field test results, key survey-design insights and resulting design decisions.

The primary goal of pretesting is to develop a questionnaire, and decision scenarios within the questionnaire, that are understandable and credible to respondents through a balanced and effective presentation of information. Survey designers desire valid and reliable value estimates, but also seek to avoid respondent fatigue from the provision of unnecessary details (Mitchell and Carson 1989; Bateman et al. 2002; Champ et al. 2017). Pretesting allows one to develop a questionnaire that can be administered to individuals with different backgrounds, interests, experiences, and knowledge levels. It provides insight into whether and how respondents understand the baseline, mechanisms for change, the change to be valued, the

payment vehicle and other components of the valuation scenario, and the broader questionnaire. This insight is necessary to ensure that respondents understand scenarios as intended. Pretesting also allows consideration of the effectiveness and balance of information presentation (e.g., using text, maps, photos, diagrams, etc.), and commonly solicits complementary input from experts (e.g., survey designers, natural scientists, communication experts, and others) to aid in the development of credible, accurate, relevant and agreed-upon scenarios.

Two types of pretesting are recommended (e.g., Mitchell and Carson 1989; Arrow et al. 1993; Bateman et al. 2002; Champ et al. 2017): (1) *qualitative pretesting* of survey materials using focus groups, cognitive interviews or other small-group methods and (2) *quantitative pretesting* using pilot studies. Qualitative pretesting can provide in-depth insights about potential subjects' comprehension of survey materials, but generally does not support statistical analyses. Quantitative pretesting enables limited statistical analyses of a pilot sample of data to test initial hypotheses, facilitate design modifications, and evaluate reliability and validity. Both types of pretesting provide distinct and complementary insights for the design of a questionnaire, and an ideal survey development process will include both approaches.[20]

Qualitative pretesting takes several different forms. Initial pretesting may use open-ended questions to gain insight on subjects' knowledge and perceptions of the change being valued. Subsequent pretests focus on specific aspects of a draft questionnaire, with final pretests allowing for initial administrations of a fully designed questionnaire to small groups of respondents. Considerable information is available on methods for qualitative pretesting (Mitchell and Carson 1989; Groves et al. 2009; Willis 2005; Krosnick and Presser 2010; Coast et

[20] Some have also suggested post-survey qualitative research with respondents to gather additional insight into respondents' understanding of the survey, interpretation of responses, and the suitability of responses to inform decision-making (Brouwer et al. 1999).

al. 2012; Champ et al. 2017). Specific studies highlight the importance of focus groups (Coast, McDonald, and Baker 2004; Desvousges et al. 1984; Desvousges and Smith 1988; Desvousges and Frey 1989; Johnston et al. 1995; Brouwer et al. 1999; Chilton and Hutchinson 1999), cognitive interviews (Kaplowitz and Hoehn 2001; Kaplowitz, Lupi and Hoehn 2004), mixed methods (Powe 2007; Pitchforth et al. 2008), and verbal protocols (Schkade and Payne 1994; Ryan, Watson, and Entwistle 2009).

Focus groups provide an efficient method for discussion of concepts and language, helping to clarify scenario descriptions, and assessing the amount and type of information that respondents require to answer the valuation question(s) (Desvousges and Smith 1988; Johnston et al. 1995).[21] However, focus groups have been criticized for the lack of independence of individuals' responses and other group-based effects (Chilton and Hutchinson 1999; Lunt 1999; Jorgensen 1999). One-on-one cognitive interviews eliminate the effects of group dynamics and allow in-depth exploration of specific design issues with individuals to an extent that is not possible in focus groups (Kaplowitz and Hoehn 2001). Hence, cognitive interviews are useful for sensitive topics or where one-on-one discussions are needed to address challenging survey design issues. However, one-on-one interviews preclude the possibility of design insights that might be revealed through group discussions. Group-based surveys in which individuals complete draft questionnaires on their own, followed by a discussion, can provide information on survey design both at the individual level and at a group level (Powe 2007).

Interviews of, or peer-reviews by, other scientists are also recommended. Peer or expert reviews (Groves et al. 2009) provide important insights into the ability of the survey process to meet the intended goal(s) of the study, based on the research experience of these experts.

[21] Guides to general focus group methods include Morgan (1997) and Krueger and Casey (2015).

Quantitative pretesting using a smaller sample from the target population (i.e. a "field pilot") is particularly important for large or high-stakes surveys where aggregate values may be large and value estimates may be controversial (Bateman et al. 2002; Champ et al. 2017). This recommendation follows similar guidance for survey research in general (Groves et al. 2009; Dillman et al. 2014). Although time and budget constraints can preclude the use of formal field pilots for some SP studies, these pilots can provide insights that cannot be derived from qualitative pretesting alone. Among the advantages of field pilot testing is the ability to assess the potential survey response rate, item nonresponse rates, the suitability and refinement of the experimental design, and bid levels for the valuation question, and to conduct preliminary investigations of hypotheses (e.g., Carlsson and Martinsson 2003; Carson et al. 2003; Champ et al. 2003; Scarpa, Campbell, and Hutchinson 2007; Vermeulen et al. 2011). Pilot testing can also include detailed post-administration debriefing of respondents and non-respondents to evaluate specific design elements, and can provide useful insights into the effectiveness of the survey design and administration process (Bateman et al. 2002; Champ et al. 2017).

Pretesting should be conducted with a sample of respondents drawn from the target population for the main study whenever possible. Pretests using non-representative groups (e.g., students in classes) may provide some information, but are not guaranteed to provide insight into performance of the survey with the target population. While it is not possible to ensure that small-group administration for qualitative pretesting will have representative samples, and qualitative pretesting often focuses on subsets of potential respondents, researchers should seek qualitative input from a combined sample of respondents that spans the diversity of respondent types that exist in the target population (i.e., the sample frame). Final field pretests should draw probability samples of respondents from the sample frame for the full survey implementation.

Pretesting should be documented, including types of pretesting, number and characteristics of respondents, and scripts used to conduct the pretests. For quantitative pretesting, the survey instruments and data should be maintained and documented. Audio and/or video recording may be useful when the substance of group or interview content needs to be reviewed at a later date. Records should be maintained of the key decisions made based at each step in the pretesting process, and published results should include clear descriptions of the survey design processes that were used. Understanding issues identified and decisions made to address issues in the survey design phase are fundamental to the content validity of any survey.

4.3. Attribute versus Non-Attribute Approaches

The validity, reliability and applicability of SP studies depend on the explanation of the change to be valued. Although there has been an increase in the use of CEs in recent years, it is not clear whether CV or CEs offer a superior approach to value elicitation in general. Each has advantages and disadvantages. For example, an appealing feature of CEs is the ability to estimate marginal values for each attribute in the study design.[22] However, this feature comes with offsetting disadvantages, such as greater complexity and the potential loss of incentive compatibility (depending on CE structure; Vossler et al. 2012). In contrast, CV offers opportunity to estimate values when an item cannot be easily defined in terms of attributes.

RECOMMENDATION 3: The use of CV or a CE to describe the change being valued should be based on how respondents tend to perceive the good, the study objectives and the information content of valuation scenarios. The processes for determining whether an attribute-based method

[22] In concept, similar information could be provided by CV, but this would require a large number of questions that each separately elicits WTP for a single permuted attribute. This is less common in the literature.

is appropriate (or not) should be clearly documented based on the change being valued and insights from survey pretesting.

The choice between CV and CEs is complex and should be based on respondent perceptions of the change being valued, the decision objective being considered, and the type of information required. For example, while many valuation applications can be conceptualized and communicated as a bundle of attributes, survey respondents may or may not think of some changes in terms of attributes. On the other hand, CEs can reveal information on the value of individual attributes that is desired to support decision-making. These marginal values of individual attributes remain invisible in a classical CV study.

The two approaches also have different advantages and disadvantages in terms of information provision, at least as commonly designed. Attributes presented using short tabular descriptions (common in CEs) can sometimes oversimplify important features of the scenario. In contrast, respondents may have difficulty identifying and distinguishing effects described using textual narratives common in CV (Hoehn, Lupi, and Kaplowitz 2010). Both approaches, depending on designs, can present complex information. Given considerations such as these, the choice of a question format should not default to attribute or non-attribute methods based solely on factors such as ease of application or prevalence in the literature.

Three primary considerations are suggested when making the decision to apply a CV or CE format. *First, will the change to be valued affect specific characteristics of the item or the item as a whole, and what are the information needs of decision-makers?* In some cases, such as estimating oil spill damages, the question may relate to estimating a value for a specified and fixed set of changes as a whole; CV supports such a decision-making context. In other applications, where a change may affect some attributes and not others, or where a range of

different attribute changes may need to be evaluated, a CE may be more applicable because it provides marginal values for individual attributes of interest, over a range of possible changes.[23]

Second, do respondents think of (and value) the change in terms of individual attributes, or as a whole? For example, if respondents think of a landscape or ecosystem holistically, then attribute framing might be inconsistent with this perspective of the change being valued. In cases such as this, parsing of the whole into its component attributes or characteristics might not fully capture the comprehensive value that respondents hold or may mischaracterize the way in which they understand improvements (Madureira, Nunes, and Santos 2005). On the other hand, individuals may think of some types of changes (e.g., to recreation sites) in terms of features consistent with a tabular, attribute-based presentation (Hoehn et al. 2010).

A related consideration concerns the underlying preference structure. Although respondents may think of the change to be valued in terms of attributes, the structure of preferences for attributes may be more complex than the linear-and-additively-separable preferences that may be implied by the matrix presentation of CE questions and is typically imposed as an assumption in basic econometric analyses of response data. In such cases, consideration should be given to whether CEs are sufficiently nuanced to permit estimation of the ways that different attributes enter respondents' utility, such that valid implicit prices (i.e., marginal WTP for each attribute) can be estimated.

Third, how does the information presentation format affect respondents' understanding of the item to be valued? If the item is large and complex, it may be difficult to describe the change to be valued adequately using a parsimonious set of attributes. In such applications, the

[23] This can be particularly useful when there is uncertainty about the physical / biological impact that might occur. In such cases, valuation of a range of possible outcomes (through different attribute levels in the CE) provides a way to ensure that the true impact has been valued or that values for a range of impacts are available to decision-makers.

use of attributes to communicate changes to be valued can increase scenario complexity, particularly when the number of attributes is large or simple attribute descriptors do not apply (Arentze et al. 2003; Hensher 2006b; Islam, Louviere, and Burke 2007; Balcombe and Fraser 2011; Dellaert, Donkers, and Van Soest 2012; Burton and Rigby 2012; Alemu et al. 2013). Complex choices can also trigger respondents to engage in simplifying heuristics or response strategies not consistent with fully compensatory, utility maximizing decisions (Kahneman, Slovic and Tversky 1982; Mazzotta and Opaluch 1995; Gigerenzer and Todd 1999; Swait and Adamowicz 2001a, b; Hensher 2006a,b; Boxall, Adamowicz, and Moon 2009; McFadden 2014; Meyerhoff, Oehlmann, and Weller 2015; Olsen and Meyerhoff 2016).[24] Challenges due to complexity can affect CV applications as well (e.g., causing subjects to apply simplifying heuristics), albeit in different and sometimes more difficult-to-observe ways (Hoehn et al. 2010).

Although this section emphasizes differences between CV and CEs, it is important to recognize that these approaches are close cousins in the genre of SP valuation and can in principle be used to estimate equivalent values. Multiple studies have investigated the convergent validity of estimates from these two formats (e.g., Adamowicz et al. 1998; Hanley et al. 1998; Cameron et al. 2002; Foster and Mourato 2003; Ryan 2004; Jin, Wang, and Ran 2006; Mogas, Riera, and Bennett 2006; Goldberg and Rosen 2007). These studies provide mixed results, but the many differences between the framing of CV and CE questions (e.g., text versus tabular presentation; Tourangeau, Couper, and Conrad 2004) can make it difficult to conduct clean and controlled comparisons. The difficulty of isolating the effects of individual CV versus

[24] Systematic conjectures (Hensher and Rose 2009; Hensher and Layton 2010), selective inattention (Scarpa et al. 2009; Balcombe, Burton, and Rigby 2011; Balcombe, Fraser, and McSorley 2015) and sequencing effects are some of the potential anomalies that may arise as respondents attempt to deal with challenging levels of complexity (Day et al. 2012). The prevalence of anomalies such as these may depend on various factors, including respondents' prior knowledge (Sandorf et al. 2016).

CE design elements when comparing these approaches speaks to the importance of testing alternative elicitation formats as a whole. The use of focus groups, cognitive interviews or other forms of qualitative pretesting can aid in the choice of an attribute or non-attribute design (Coast 1999; Bennett and Blamey 2001; Coast and Horrocks 2007; Coast et al. 2012; Riera et al. 2012).

4.4. Experimental Design

We discuss experimental design in the context of bid assignment in discrete-choice CV and attribute-level assignment in CEs. Experimental design in these contexts, and particularly for CEs, is complex and evolving (Sándor and Wedel 2002; Ferrini and Scarpa 2007; Scarpa and Rose 2008; Rose et al. 2008, Rose and Bliemer 2009 2014; Bliemer and Rose 2010; Vermeulen et al. 2011; Johnson et al. 2006, 2013). It defines the manner in which different treatments (bids or attribute levels) are assigned to a question (or sequence of questions) within a questionnaire and how blocks of questions are assigned to different survey versions. It can be as simple as developing the bid structure for a single binary-choice question or as complex as a choice question with multiple attributes, multiple alternatives within a question, multiple questions within a questionnaire, and multiple versions of the survey instrument (Caussade et al. 2005).

RECOMMENDATION 4: The primary goal of experimental design in CV and CEs is to develop designs that yield efficient and unbiased estimates of preference parameters and value estimates. Designs should make use of information from prior empirical research and require pretesting. All treatment effects and relevant attribute interactions should be individually or jointly identified. Experimental design should generally allow for interactions (and perhaps other types of non-linear-in-attributes utility functions), consider both statistical efficiency and respondents' cognitive abilities and attention budgets, employ constraints on implausible attribute levels and

combinations, use designs that are robust to alternative model specifications, and consider how the levels chosen for each attribute influence design properties. SP studies should report sources of *a priori* information used to formulate designs, provide evidence to support the design chosen, and report the steps in the design process.

For a CV question, the investigator must choose cost or bid amounts with consideration for range and spacing. Effective designs for these questions ensure that monetary amounts are credible to respondents, enable unbiased welfare estimates, and minimize the variance of these estimates (Cooper and Loomis 1992; Cooper 1993; Kanninen 1993a,b, 1995; Alberini 1995; Haab and McConnell 2002). For CEs, basic design considerations involve four related components: (1) selecting attributes and levels for each attribute, (2) deciding how many alternatives will be in each question and the number of questions each subject will answer, (3) determining how attribute levels will be combined to form different alternatives, and (4) determining how the question/attribute combinations will be blocked for presentation to subsets of respondents.[25] Effective multi-attribute designs enable statistically efficient welfare estimates that are robust to specification of the model and ensure identification of important attribute main effects and attribute interaction effects.

Experimental design decisions are informed by multiple factors including the change to be valued, prior studies on the specific application, and insights learned through the survey design process. Attributes and levels should be selected based on a combination of the values needed to support decision-making, feasibility of implementation, plausibility to respondents, and statistical efficiency (Johnson et al. 2013). Levels must also be chosen in light of the

[25] Question/attribute combinations can also be randomly assigned to individual respondents, although this may result in a forgone opportunity to increase estimation efficiency.

functional form(s) to be used for utility. For example, one might expect linear rates of change along some utility dimensions or non-linear changes along others. The number of levels for each attribute must be sufficient to model these effects.

Statistical design properties (e.g., D-efficiency, D-optimality, C-efficiency, S-efficiency) are derived based on an assumed statistical model and set of intended hypothesis tests (Scarpa and Rose 2008; Kessels, Goos, and Vandebroek 2006). At a basic level, designs can follow guidance such as that provided by Kuhfeld (2005) or develop designs using software such as Ngene (http://www.choice-metrics.com/features.html, accessed June 2, 2016) or SAS (https://support.sas.com/rnd/app/qc/qc/qcdesign.html, accessed June 2, 2016), although it is important to understand the implications of any programmed design algorithm. There is no single best design for all purposes; designs should anticipate the primary estimating specifications, targeted estimates, and hypotheses to be tested (Scarpa and Rose 2008; Rose and Bliemer 2009, 2014; Bliemer and Rose 2010; Kerr and Sharp 2010; Johnson et al. 2013). For example, assuming a parametric model to analyze the effect of bid levels on the probability of a "yes" response to a binary-choice CV question may suggest a specific allocation of bid amounts, but the optimal (or efficient) bid design might be different under alternative model assumptions (Kanninen 1993a,b, 1995; Alberini 1995; Scarpa and Bateman 2000). Moreover, the true data-generating process (and thus the best statistical model) is inevitably unknown. Thus, experimental design, as in any statistical optimization process, is a form of 'chicken and egg' problem. Design uncertainty can be reduced using information obtained from previous work and pilot studies, and Bayesian design methods may be applied to formally reflect uncertainty in expected parameter values (Kessels et al. 2009; Rose, Scarpa, and Bliemer 2009).

In some cases, complex designs may not lead to statistical improvements, for example

due to tradeoffs between statistical efficiency and respondents' cognitive capacity (or response efficiency). Experimental design for an SP study is not simply a mechanical process in which a programmed optimization routine can always provide the best design. Designs should hence be subjected to qualitative pretesting to consider how respondents react to the offered mixes of attributes, and pilot studies (where possible) should be used to calibrate statistically determined designs and determine the reasonableness of the design priors (e.g., any assumed parameter values based upon which the optimal design has been deduced).

Other issues that should be considered when developing an experimental design include, but are not limited to, information order effects (Chrzan 1994; Kjaer et al. 2006), attribute non-attendance (Scarpa et al. 2009; Campbell et al. 2008a; Scarpa, Thiene, and Hensher 2010; Campbell, Hensher, and Scarpa 2011; Hole 2011; Hensher, Rose, and Greene 2012; Hole, Kolstad, and Gyrd-Hansen 2013), omitted attributes (Petrin and Train 2003), bid amount effects (Boyle, Johnson, and McCollum 1997; Hanley, Adamowicz, and Wright 2005), statistical power (Vossler 2016), and effects of the chosen optimization criteria (Yao et al. 2015; Olsen and Meyerhoff 2016). As discussed above, respondents may develop coping strategies (heuristics) to deal with choice complexity, suggesting that designs should include a limited number of attributes that are particularly relevant to decision-makers and respondents. Louviere et al. (2008) and DeShazo and Fermo (2002) find that as statistical efficiency and choice complexity increase, the consistency of respondent choices declines. These and other findings highlight the need to balance design efficiency against respondents' cognitive capacities (as well as their opportunities and inclinations to digest all aspects of choice sets). For these and other reasons, *ex ante* pretesting, (discussed above) and *ex post* robustness checks in data analyses (discussed below) are important to establish the credibility of empirical results associated with any

experimental design.

4.5. Ethics in Data Collection

SP methods are increasingly applied to inform public decisions, placing heightened awareness on issues of research ethics. Familiarity and compliance with recognized standards for research ethics and the protection of human subjects can help ensure that SP research is above reproach.

RECOMMENDATION 5: SP protocols should be reviewed by university or other review boards tasked with protection of human subjects, and informed consent should be obtained from subjects. The rights and welfare of human subjects should be a paramount concern in all aspects of a SP study, including research implementation and reporting. Survey procedures should avoid deception that may have significant negative consequences for respondents, unintended influences on study outcomes or validity, or that compromise the ability to use study results to support decision-making.

SP research requires interaction with human subjects. In most cases, those conducting a SP survey are required to follow established procedures for the protection of human subjects in research (e.g., Common Rule in the US; Tri-Council standards in Canada; Governance Arrangements for Research Ethics Committees in the UK). These procedures are based on principles of respect, informed consent, minimizing risks to the subjects, and fairness. For example, survey research conducted by US academic institutions (including questionnaires, focus groups and interviews) is subject to Institutional Review Board (IRB) oversight, with a few exceptions. Within the UK, the requirements for ethical review by Research Ethics Committees are set out in the UK-wide edition of the Governance Arrangements for Research Ethics

Committees (GAfREC), published by the UK Health Departments in 2011. Additional guidance is provided by groups like the European Society for Opinion and Marketing Research (ESOMAR) and the World Association for Public Opinion Research (WAPOR), and funding bodies such as the Economic and Social Research Council and the Medical Research Council in the UK, and the National Science Foundation in the US. [26] Codes of ethics such as these are also relevant to the design, implementation and reporting of SP studies. Requirements may vary by the context of the research and across countries and institutions, but all SP studies should adhere to at least minimum standards with regard to the rights and protections of human subjects.

Research ethics issues in SP have parallels to those in experimental economics, where there has been debate about the benefits of research relative to the risks associated with lack of informed consent, breaches of confidentiality, deception, and other concerns (e.g., Riach and Rich 2004; List 2011, 2009). Among the concerns relevant to SP research is the use of deception.[27] This is a complex issue. Some types of deception in economics research might be considered innocuous, and deception may be necessary to answer certain research questions (Rousu et al. 2015). However, deception can have negative consequences. For example, presentation of inaccurate information may lead to behavior change that could harm respondents, diminish the perceived validity of study results, or affect future research.[28] Ethics guidance does

[26] The American Association for Public Opinion Research provides a similar code of ethics for the collection of primary data (http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx, accessed August 4, 2016). See http://www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics/, http://www.mrc.ac.uk/research/policies-and-resources-for-mrc-researchers/good-research-practice/, and https://www.nsf.gov/bfa/dias/policy/rcr.jsp, all accessed October 4, 2016

[27] For example, passive deception may occur if respondents are led to believe that elements of the valuation scenario are actual options being considered when this is not the case, or when respondents receive different treatments or scenarios without explanation.

[28] For example, inaccurate presentation of health risks could lead subjects to over/under-invest in protective actions in real life. Presented baselines can also influence behavior (Whittington 2004). Other potential consequences include nonresponse, protests, scenario rejection, refusal to participate in future research or contamination of the subject pool (Croson 2003; McDaniel and Starmer 2003; Jamison, Karlan, and Schechter 2008; Rousu et al. 2015).

not preclude all deception, but the risks must be evaluated relative to the benefits. This balance will differ across contexts (Morrison 2002). Rousu et al. (2015) and Colson et al. (2016) suggest that clear definitions of deception and identification/avoidance of harmful deception can help to determine and reduce the extent to which specific practices violate the norms of research ethics.

Survey implementation methods, especially in developing countries, may also be implicitly or explicitly coercive (Whittington 2004). For example, consider a country where citizens may fear their government. If a survey is endorsed by a government agency, or if enumerators appear to be from a government agency, respondents may feel pressured to respond. Issues such as these should be considered in light of the particular context where the study is being conducted and should be evaluated during survey pretesting.

Other ethical concerns relate to confidentiality, data storage and study reporting. For example, questionnaires may collect sensitive information (e.g., data on health status or income), the release of which may have negative ramifications for individuals or groups. Hence, SP research should engage in rigorous practices to protect subject confidentiality (this is generally required by ethics boards such as IRBs). Standard practices include the use of pre-specified protocols for data coding, storage, protection, access and disposition. Ethics related to study reporting also extend beyond human subjects concerns. For example, study reporting (e.g., in publications or online appendices) should enable others to understand, evaluate the merits of, and replicate the methods that were applied (Vossler 2016).

Given such concerns, processes to obtain informed consent and to reduce the chance of harm to research subjects should be incorporated into all study designs.[29] A particular concern

[29] Consent procedures need not require the return of a signed consent form. For example, consent information can be included in a survey cover letter, with an explicit statement that participation is voluntary and returning the survey

arises when SP data are collected by entities not subject to IRB or similar oversight. The Editorial Expression of Concern published by the editors of the Proceedings of the National Academy of Sciences highlights some of the issues associated with data collected outside academic institutions where such oversight may not be present (Verma 2014). In these instances, guidance should be sought on acceptable ethics review and practice.

4.6. Extent of the Market, Survey Mode, Sampling, and Non-response Bias

SP research relies on multiple implementation modes to collect data and sampling methods to identify respondents, including internet survey administration and sample compilation methods not foreseen at the time of the NOAA Panel (Arrow et al. 1993). Common modes include traditional mail and in-person surveys, along with newer approaches such as internet administration, mixed-mode surveys and other electronic methods (tablets, etc.), and valuation workshops (MacMillan et al. 2002). Each mode has advantages and disadvantages, and there is no consensus around a single best method (for a general discussion, see Dillman et al. 2014).[30]

RECOMMENDATION 6: The most appropriate mode of data collection is context specific and the rationale for the selected mode should be documented. There are advantages and disadvantages to all survey administration modes. However, given the inability to effectively convey complex valuation materials in a telephone interview, this survey mode should be used with caution. Samples should be drawn from known frames that are consistent with the population for which values are to be estimated, and respondents should be randomly selected

implies consent to participate in the research. Regardless of how consent is obtained, participation should be voluntary and respondents should be free to stop participation at any time and refuse to answer any or all questions.

[30] For example, the use of personal interviews (the mode recommended by the NOAA Panel), while allowing trained interviewers to guide the interview administration process, can be subject to social desirability and other interviewer effects (Boyle and Bishop 1988; Andreoni 1989; Blamey et al.1999; Leggett et al. 2003).

from the sample frame using an explicit sampling procedure. Contemporary approaches should be used to identify and mitigate non-response bias, including survey-design features and the collection of data to aid in identifying and characterizing non-response patterns. Whenever possible, analysts should not rely solely on response rates and demographic data to infer the presence or absence of non-response bias.

Survey mode and sampling considerations affect, and are affected by, multiple aspects of survey design, including sample frame and representativeness.[31] Recent research suggests that data collection mode does not substantially influence SP study outcomes, although the results are mixed and context specific (Lindhjem and Navrud 2011a, b; Windle and Rolfe 2011; Bell, Huber, and Viscusi 2011; Ščasný and Alberini 2012; Boyle et al. 2016; Sandorf et al. 2016).

The Extent of the Market (Affected Population) and Sample Frame

The extent of the market is in principle defined as the group of people whose welfare would be affected by the change being valued in a SP study, or an affected population. While the extent of the market may be easily identified for some types of use values, it can be difficult to identify for non-use values. Moreover, differences between geo-political jurisdictions and the locations of those who are affected can influence our understanding of the admissible extent of the market. For example, for many national policy decisions, residents of other countries affected by the

[31] A mail survey requires a complete list of names/addresses for eligible respondents. Web survey respondents are often recruited to an internet panel and then asked to complete surveys once they become members of the panel. Assuming the initial lists from which the samples are selected are representative of the target population, these two approaches have different impediments to response. For a mail survey, respondents can see the topic and then decide whether to reply, introducing possibilities of topic-related selection biases. For panel-based web surveys, the panel recruitment process can influence the pool from which respondents are drawn, and respondents who agree to be on such panels may not reflect the population across all observable and unobservable characteristics. With phone administration, sample recruitment can be accomplished with random-digit dialing (RDD), but the proliferation of cell phones, caller ID and number blocking has complicated sampling and reduced response rates. Lines per household may vary by income and other factors, affecting the probability that any household will be reached.

policy typically have no legal standing in a formal benefit-cost analysis. These individuals are hence excluded from the market and sample frame (Loomis 1996).

Determining the extent of the market is important to ensure that an appropriate sampling frame is selected. It is also necessary when expanding (or aggregating) individual value estimates to population values (i.e., calculating aggregate benefits or costs). A clearly defined market area further facilitates determinations of whether a sample frame is available and complete (i.e., whether all affected individuals have a known probability of selection into the sample). An identified sample frame is fundamental to the identification of nonresponse effects.

The true extent of the market is typically unknown. Hence, geopolitical boundaries or features of the change to be valued (e.g., watershed areas affected) are often used to identify the market and associated sample frames. However, some research has considered empirical approaches to identify the market extent or conducted sensitivity analysis over different market definitions (Loomis 1996, 2000; Vajjhala, John, and Evans 2008; Sanchirico et al. 2013; Morrison 2000; Johnson et al. 2001). Other work has specified the value estimate as a function of distance from the affected area (Hanley et al. 2003; Banzhaf et al. 2006; Bateman et al. 2006a; Schaafsma, Brouwer, and Rose 2012; Schaafsma et al. 2013), but such "distance decay" is not always revealed and may not apply for all goods (Boxall et al. 2012; Johnston and Ramachandran 2014; Johnston et al. 2015). Thus, the determination of the extent of the market, while a critical component in selecting a sample frame for an SP study, remains a challenge for which research is warranted.

Survey Mode and Sampling

In-person interviews have the desirable feature that the survey completion process can be guided

by a trained interviewer, but this approach is the most expensive data collection format and survey responses may be affected by unintended interviewer effects. Telephone surveys are relatively inexpensive and are convenient for quick data collection, but have experienced declining response rates (Brick and Williams 2013), and have been forced to account for the growing proportion of the population that can be reached only on a cell phone (e.g., Blumberg et al. 2012).[32] Many countries provide access to representative sample frames (i.e., address lists) of households for mail surveys. However, the problem of sampling individuals randomly from within households remains a challenge (Link et al. 2008). A recent development has been the adoption of the internet for survey data collection (Tourangeau, Conrad, and Cooper 2013). Most internet surveys are based on *opt-in* panels—panel members are volunteers who have, most commonly, been recruited on-line. A sample is selected from among panel members and invitations are sent asking recipients to complete specific surveys. The sample may be selected using quotas that match to population benchmarks for demographic characteristics. Still, samples from opt-in panels are not probability samples from the general population, and may be subject to unknown selection biases related to people's unobserved characteristics (Baker et al. 2013).

High-quality surveys tend to use addresses as a frame for mail and web surveys of the general population. When using address-based sampling (ABS) for web surveys, invitations are sent by mail to a sample of addresses asking residents to complete an on-line questionnaire. The same sequence of contacts long advocated by Dillman (1978, 1991) for mail surveys can be used to invite sample members to respond via the web (e.g., Messer and Dillman 2011; Millar and Dillman 2011; Tancreto et al. 2012). For telephone surveys, dual-frame samples including both landline and cell numbers provide the best coverage (Lohr and Brick 2014).

[32] Moreover, cell phone numbers cannot be as readily assigned to specific geographic areas, making it harder to match the respondent to their neighborhood's characteristics.

Concerns regarding the use of the internet-based methods for surveys revolve around issues of representation of the population given varying levels of computer-literacy across age groups, varying degrees of connectivity to the internet across different sociodemographic or economic groups, and the extent to which respondents "take care" when answering the questionnaire. Lindhjem and Navrud (2011b) reviewed 17 SP studies in which comparisons between internet and other survey formats are made, and found only minor differences in values (see also Boyle et al. 2016 and Sandorf et al. 2016). In some circumstances, however, the use of the internet survey format remains problematic. For instance, in developing-country or other contexts where literacy rates are low and household internet penetration is weak, personal interviews may be the only practical format (Whittington 1998; Bennett and Birol 2010).

Response Rates and Nonresponse Bias

Most high-quality surveys continue to apply established methods to increase response rates, although response rates are a poor indicator of nonresponse *bias* (Groves 2006; National Research Council 2013).[33] Methods include participation incentives (Singer and Ye 2014; Mercer et al. 2015); advance letters (de Leeuw et al. 2007); multiple contact attempts (Dillman et al. 2014); two-phase sampling (Groves and Heeringa 2007); refusal "conversion" (persuading reluctant respondents to take part despite their objections); and longer field periods (which allows for more contact attempts and better spacing). Economists often worry about how explicit and implicit incentives might affect SP estimates, but little work has examined this issue.

Ideally, concerns about survey nonresponse will be addressed *ex ante* using effective

[33] Many studies have found a weak relationship between response rates and nonresponse bias (Curtin, Presser, and Singer 2000; Keeter et al. 2000; Merkle and Edelman 2002; Keeter et al. 2006; Groves and Peytcheva 2008), and efforts to increase response rates can sometimes worsen nonresponse bias (Peytchev 2009; Lundquist and Särndal. 2013).

survey design and administration. However, selection biases may persist even with attention to such issues, and the resulting biases can be a concern (Edwards and Anderson 1987; Whitehead 1991; Whitehead, Groothius, and Blomquist 1993, 1994; Messonnier et al. 2000; Krupnick and Evans 2008). *Ex post* approaches to investigate nonresponse bias are increasingly recommended by the survey literature, and are discussed in the section addressing SP data analysis below.[34]

5. VALUE ELICITATION

Methodological choices related to value elicitation can be split into five general categories: (1) whether attribute or non-attribute methods are applied; (2) the type of welfare measure elicited; (3) the framing of response options for the chosen question format; (4) the choice of payment vehicle; and (5) the use of auxiliary questions and other design elements to support and evaluate validity. The first of these is addressed under the discussion of survey design above. Categories (2)-(5) are addressed here. We discuss issues such as incentive properties and other elements of value elicitation required to ensure valid and reliable welfare estimation. These topics are addressed in more detail in a separate section later in the paper. The choice of approach depends on the valuation context in question, including the type of good, the nature of the decision being evaluated and associated institutions (property rights, tax system, etc.).

Concerns of complexity and incentive compatibility cut across all of these choices. SP survey design must balance the information required to support decision-making and preference elicitation, while ensuring that the valuation exercise is plausible and not overly complex. An

[34] Those administering SP surveys should also be concerned about item nonresponse to specific questions within a survey, which is not explicitly addressed here. A field pretest can be used to identify questions that may be subject to item nonresponse so that these issues can be resolved prior to full survey implementation. For example, the common practice of asking for income later in a survey and using categories for respondents to report income are two examples of practices that have been developed to minimize item nonresponse to the income question.

overall goal should be to present respondents with an incentive compatible valuation exercise that involves a plausibly consequential decision. Such designs minimize the opportunity for strategic and other types of overt, as well as inadvertent, anomalous response behaviors by respondents. The literature demonstrates that for public goods, ideal designs include choices where the item being valued is clearly understood, payment is binding if the proposed change is put into practice, respondents perceive their responses as influencing the provision of the item being valued (i.e., consequentiality applies), and other aspects of the elicitation format (e.g., number of questions and alternatives per question) encourage truthful preference revelation.

5.1. Willingness to Pay versus Willingness to Accept

Kim et al. (2015) review the literature addressing differences between willingness to pay (WTP) and willingness to accept (WTA) estimates, considering theoretical and empirical explanations. Their review illustrates multiple reasons why parallel WTP and WTA estimates can diverge, and multiple possible relationships between these empirical estimates and underlying Hicksian welfare measures. From conceptual and theoretical perspectives, implied property rights (not necessarily legal entitlements, see Knetsch 2007) of the change under consideration, and the respondent's logical reference condition, should inform the choice of welfare measure (Carson, Flores, and Meade 2001; Kim et al. 2015). At the same time, established empirical difficulties associated with WTA estimation lead most studies to estimate WTP.[35]

RECOMMENDATION 7: The decision context generally determines whether WTP or WTA is the most appropriate welfare measure from a conceptual perspective, but the final choice of

[35] Some recent research has attempted to address some of these empirical challenges, for example via the use of provision point elicitation mechanisms (Bush et al. 2013).

welfare measure should be motivated by a combination of theory and empirical considerations, and the motivation for this choice should be explained. Given that WTA estimation often faces practical challenges, such as difficulty in framing incentive compatible questions and increased rates of scenario rejection, WTP estimation will often lead to estimates with superior empirical properties, but may overestimate or underestimate the true welfare measure if WTA is conceptually appropriate. However, WTP estimation should not always be considered the default, and WTA estimation should be applied when it is appropriate and feasible.

The NOAA Panel (Arrow et al. 1993) recommended WTP estimation. However, with advances since then—both in our understanding of why WTP and WTA might diverge (Brown and Gregory 1999; Horowitz, McConnell, and Murphy 2013; Kim et al. 2015; Ericson and Fuster 2014) and in the design of SP methods—this recommendation need not always hold. In cases where payment reductions[36] are institutionally feasible and incentive-compatible questions can be designed, WTA estimation may be practical and appropriate. The WTP versus WTA choice can be made using an understanding of the literature on this topic, supplemented by information from pretesting that provides insight into the considerations outlined by Kim et al. (2015).

5.2. Valuation Question Response Formats

There are multiple response formats for CV and CEs (Mitchell and Carson 1989; Adamowicz et al. 1998; Carson et al. 2001; Hanley et al. 2001; Bateman et al. 2002; Carson and Louviere

[36] Examples include tax or fee reductions associated with environmental management changes (Johnston and Swallow 1999) or the use of financial incentives to change health-related behaviors (Promberger, Dolan, and Marteau 2012).

2011).[37] Within each of these formats, each subject may be asked to consider one or multiple valuation questions. There are advantages and disadvantages of each response format, including differences in incentive properties (Carson and Groves 2007; Carson et al. 2009; Collins and Vossler 2009; Harrison 2010; Carson et al. 2014). Although research in this area is evolving and new information on response formats continues to emerge (e.g., Vossler et al. 2012; Vossler and Holladay 2016), the literature offers some guidance to inform the choice of response format.

RECOMMENDATION 8: Incentive compatible response formats are preferred for the valuation of public goods. Of the currently available formats for CV and CEs, the most straightforward means to achieve incentive compatibility is through the use of a single binary-choice question for each respondent, generally (but not always) consisting of a baseline or status quo alternative versus the change being evaluated. If a format with less clear incentive properties is applied, the reasons for choosing such an alternative format should be explained, along with the implications for welfare estimates. If investigators ask multiple valuation questions of each subject, additional tradeoffs involving issues such as efficiency, bias, and the evolution of choice heuristics should be considered, and question order should be randomized across respondents.

Of currently available CV response formats, the use of a single, binary-choice question is preferred for public goods given the established incentive properties of this format (Carson and

[37] Question framing/response alternatives for CV include, but are not limited to (a) iterative bidding, (b) open-ended elicitation, (c) payment cards (without or with anchors), and (d) binary or dichotomous choice (single bounded, one-and-one-half bounds, double bounded or multiple bounded). CEs have included various types of binary and multinomial choice formats, along with some types of ranking and best/worst scaling (Marley and Louviere 2005; Flynn 2010; Scarpa et al. 2011), although there has been some inconsistency in the literature regarding which elicitation formats are included under the umbrella of CEs (Carson and Louviere 2011).

Groves 2007; Carson et al. 2014).[38] Other available formats violate incentive compatibility, are incentive compatible under a narrower and more restrictive set of circumstances than binary choices, or have poorly understood incentive properties.[39] For example, variants of binary-choice questions (e.g., one-and-one-half bounds, double-bounded and other multiple-bounded framing), while enhancing the statistical efficiency of estimation (Hanemann, Loomis, and Kanninen 1991; Cooper and Hanemann 1994; Cooper, Hanemann, and Signorello 2002), generally violate incentive compatibility.[40] The use of open-ended questions has decreased in recent years relative to other formats, in part due to the large number of respondents who provide either unrealistically high or zero WTP responses. This finding is consistent with a lack of incentive compatibility and attendant strategic behavior (Carson and Groves 2007). Empirical evidence on validity (discussed below) also supports the use of incentive compatible binary-choice formats.

CEs often apply binary- or multinomial-choice formats. When applied to public goods, CE questions commonly include a status quo (baseline) alternative and one or more alternatives which include changes from that status quo. As in CV applications, binary-choice CE formats[41] provide the most straightforward avenue to ensure incentive compatibility, particularly if a single valuation question is asked of each respondent (Carson and Groves 2007; Collins and Vossler 2009; Vossler et al. 2012). Multinomial choice formats can also be incentive compatible, albeit under a more restrictive set of conditions than binary choices (Carson and Groves 2007; Collins

[38] The conditions under which a binary-choice question is incentive compatible for public goods are well-established, particularly when questionnaires include a single choice question. These conditions include a choice perceived to be consequential and a binding payment mechanism (Carson and Groves 2007).

[39] Recent work suggests possible avenues for other formats. For example, Vossler and Holladay (2016) identify assumptions under which open-ended and payment card formats are incentive compatible. As they discuss, however, prior implementations of these are unlikely to have met these assumptions.

[40] If respondents are unaware that they will be asked multiple valuation questions, then responses to subsequent valuation questions should not affect the incentive compatibility of the initial question.

[41] A binary- or dichotomous- choice CE question asks the respondent to choose a single preferred alternative from a choice set that includes two multi-attribute alternatives (or choice options), typically involving a choice between a status quo alternative and a single non-status quo alternative.

and Vossler 2009). However, the contextual nature of CEs implies that flexibility is needed in the selection of a format. For example, binary choices may not always be a good match to real-world decisions, including choices over publicly provided goods like recreation opportunities where multiple alternatives are frequently available in real-world choice contexts. In such cases, the choice of format should balance incentive properties with realism and relevance.

Summary articles on CEs generally do not discuss or present recommendations on the number of alternatives presented to respondents (de Becker-Grob et al. 2012). However, evidence suggests that CE responses can be sensitive to the number of alternatives.[42] This evidence is mixed on whether binary- or multinomial-choice formats are preferred, further supporting the need for some flexibility in this regard, and the need for further research.

These concerns reflect a more general recommendation that incentive properties should be only one of the considerations that influence the selection of a response format. All formats have potential advantages and disadvantages, beyond those that have been linked directly to incentive properties. For example, subsequent answers to iterative-bidding questions may be susceptible to anchoring on initial bid amounts (Boyle, Bishop and Welsh 1985; Thayer 1981), while payment cards can be afflicted by range effects (Rowe, Schulze, and Breffle 1996; Covey, Loomes, and Bateman 2007; Smith 2006).[43] The single binary-choice format may be associated with unintended effects on value estimates due to bid anchoring and insufficient responsiveness to bid amounts, among other aspects of bid design (e.g., Holmes and Kramer 1995; Kanninen 1993a,b; 1995; Herriges and Shogren 1996; Boyle et al. 1997; Bateman et al. 2009).

[42] For example, Boyle and Özdemir (2009) found that giving CE respondents two versus three alternatives influenced responses. Zhang and Adamowicz (2011) suggest that this result occurs due to the offsetting effects of task complexity and preference matching that increase/decrease, respectively, the probability of choosing the status quo option. Rolfe and Bennett (2009) found that binary-choice CE formats led to greater rates of serial non-participation, compared to multinomial-choice formats.

[43] Anchoring can affect responses to other types of response formats (e.g., Green et al. 1998).

In applications with multinomial response formats, investigators should also consider the complexity and difficulty of the choices (e.g., in pretesting) to avoid experimentally induced errors (DeShazo and Fermo 2002). Swait and Adamowicz (2001a,b) found that preference coefficients change non-proportionally with task complexity in CEs. However, others have found no effect on value estimates (Meyerhoff et al. 2015), as would be the case if all of the choice-model coefficients change proportionally, equivalent to a change only in the model's error variance. Thus, while the case for moving beyond binary choices (status quo and one alternative) is application-specific, survey design and data analyses should consider the potential effects of response anomalies that may be introduced in any chosen response format.

Non-choice formats, such as rankings, have other concerns such as the provision of unanchored preference indices that do not necessarily reveal whether a respondent would pay for the change (Boyle et al. 2001; Louviere, Flynn, and Carson 2010). We do not recommend the use of these and other conjoint response formats, as such formats face multiple challenges in the context of utility-theoretic welfare estimation (Louviere et al. 2010). Ranking tasks can also be more cognitively challenging than choose-one tasks (Yan and Tourangeau 2008). Recent applications have considered best/worst ranking as an alternative (Marley and Louviere 2005; Flynn 2010; Scarpa et al. 2011).[44] This approach, like multiple-bounded CV questions, can enhance efficiency and is useful when viable sample sizes are small. However, the use of this approach for welfare analysis hinges on additional assumptions (Scarpa et al. 2011), and some of the other challenges of rankings still apply (e.g., complex incentive properties).

Concerns with alternative response formats, such as the examples discussed here, suggest

[44] There are three types of best/worst elicitation, only one of which is applicable to neoclassical welfare analysis. This approach (sometimes called "case 3") asks respondents to choose their most preferred (best) and least preferred (worst) option out of three or more multi-attribute alternatives (Flynn 2010). Each question thereby provides a more complete preference ranking than an otherwise identical choose-one DCE.

that binary/multinomial choice formats are a more dependable option for welfare analysis at the present time. However, we encourage research to evaluate the comparative properties and performance of alternative response formats such as best/worst ranking.

Number of Valuation Questions

All types of SP questionnaires can ask respondents to answer multiple value elicitation questions within the same instrument, and multiple-question formats are common in CEs. Multiple valuation questions allow for within-subject elicitation of preference information over different mixes of attributes for the item being valued and allow for rigorous testing of compliance with fundamental axioms of choice (Johnson and Matthews 2001). Multiple questions also allow increased efficiency of value estimates for any given sample size (which may be important given limited budgets for survey administration) and may provide respondents an opportunity to develop a better understanding of the task at hand (including opportunities to learn about or become familiarized with unfamiliar goods or valuation contexts). However, the assumptions required for incentive compatibility are stronger when respondents answer multiple valuation questions (Vossler et al. 2012), and responses to these questions may be subject to sequencing or anchoring effects. In addition to learning opportunities, multiple questions may lead to respondent fatigue and/or the development of simplifying choice heuristics, all of which may require additional structure in the conceptual and empirical model used to estimate values. Hence, the choice to use multiple valuation questions requires careful consideration.

There is a significant literature addressing sequence effects, when/how they occur, and implications for welfare estimation (e.g., Boyle et al. 1993; Carson and Mitchell 1995; Bateman and Langford 1997; Carson, Flores, and Hanemann 1998a; Holmes and Boyle 2005; Day and

Prades 2010; Bech, Kjaer, and Lauridsen 2011; Day et al. 2012; Craig et al. 2015). If each choice task presented to a respondent is presumed to be independent of the other choices in the sequence, then the absence of a sequencing effect is desirable. However, if later elicitation questions are constructed in a way that depends on the conditions valued in prior valuation questions, or the respondent's answers to those questions, then it is expected that values estimated from responses to later questions should be influenced by responses to earlier ones (Cameron and Quiggin 1994; Carson and Mitchell 1995; Carson 2012). Further, if respondents learn about the change being valued and their preferences evolve or converge through the sequence of questions, then answers to later questions in the sequence may better reflect actual preferences than answers to earlier questions (Holmes and Boyle 2005; Bateman et al. 2008a; Brouwer et al. 2010a; Carlsson et al. 2012).[45]

Given these and other tradeoffs, consideration should be given to whether the additional information and efficiency afforded by multiple valuation questions offsets the potential complications, including implications for sequencing,[46] less-clear incentive properties, and greater complexity burden (Swait and Adamowicz 2001a,b; Caussade et al. 2005; Vossler et al. 2012). These potential effects should be investigated in the survey design process and the robustness of value estimates should be investigated in econometric analyses of response data.

5.3. No-Answer Option

The NOAA Panel recommended that respondents be given a "no choice" or "no answer" option

[45] Research is required to consider how these responses are treated in econometric analyses (e.g., discarding or down-weighting some of the early choices as part of a "burn in" phase for the respondent).

[46] It may be possible to reduce the potential for sequencing by preparing respondents for the types of choices they will be asked to make; this is common in CEs. A "visible choice set" is said to be in effect if respondents have (a) pre-notification that they will face multiple valuation questions, and (b) some information on the range of situations they will be asked to consider (Bateman et al. 2004).

(Arrow et al. 1993). For a binary-choice question, this would allow respondents to decline to answer either yes or no; for a CE, it would allow respondents to decline to choose either the status quo or one of the alternatives provided. Such answers could arise if respondents are unable to make a decision because they (a) need additional time to think (Schuman and Presser 1979), (b) do not have an opinion, (c) equivocate, (d) are indifferent between the options, or (e) fail to understand the choice context (Feick 1989). Similar but not identical to no-answer options are polychotomous response options suggested by Blamey, Bennett, and Morrison (1999) and Loomis, Traynor, and Brown (1999). These provide a means for respondents to express support for the outcome under consideration, while simultaneously indicating that they would/could not pay for it.[47]

RECOMMENDATION 9: The preferred response format to SP questions need not always include a no-answer option that is distinct from the status quo.

A no-answer option reduces the pressure on respondents to give a definitive answer, but may also allow some respondents to avoid exerting the necessary cognitive effort to make a choice (Krosnick et al. 2002). CV studies applying binary-choice questions have found that applications including and excluding a no-answer option can yield comparable results, e.g., those who choose the "no-answer" option answer "no" when this option is excluded (Carson et al. 1998b; Groothius and Whitehead 2002; Krosnick et al. 2002). Similar research is needed for CEs. In sensitive applications, ethical considerations may suggest that respondents be given the opportunity not to answer. Tourangeau and Yan (2007) discern that more misreporting occurs when sensitive respondents are considered. Topic sensitivity may be a significant concern, for

[47] Such options can help respondents resolve competing response motivations that cannot be reconciled with a single yes/no response (Blamey et al. 1999; Loomis et al. 1999). Also see Ready, Whitehead, and Blomquist (1995).

example, in SP studies of (a) hotly contested local issues, (b) certain health conditions, (c) issues in developing-country applications where respondents may view enumerators as being in a position of power. Survey pretesting should consider the sensitivity of the survey topic and implications for a no-answer option.

5.4. Decision Rule

The decision rule described in an SP questionnaire states the relationship between responses and provision of the change under consideration. The NOAA Panel (Arrow et al. 1993) recommended that questions be framed as a referendum vote, where the implied implementation criteria would be majority rule. When considering this recommendation, we note that the Panel's specific focus was the estimation of non-use values for public goods in the US. In contrast, this article considers broader applications that may include use values and contexts in which referendum votes may be inapplicable or unrealistic. For example, a referendum may not be an appropriate decision rule in political settings where direct democracy is not practiced or when referenda are not used to determine the provision of public goods.

RECOMMENDATION 10: A decision rule should be selected that is realistic and binding on respondents. Referendum formats should be considered where plausible, but are not always relevant to the choice context in question.

Past research provides some insight into the effects of stated decision rules. Taylor et al. (2010) found that the *absence* of a clearly specified decision rule increased value estimates relative to cases with posted-price and plurality-vote decision rules, but there was no difference in value estimates between the two specific decision rules. Vossler et al. (2012) found that

truthful preference revelation required a financially binding decision rule and more than just a weak chance that answers to the valuation question would influence actual decisions. Similar findings are reported by Carson et al. (2014) and Vossler and Evans (2009).

These insights support the use of a majority vote for public good valuation where a referendum is a plausible decision mechanism. In instances where referendum votes do not ordinarily apply, such as recreation choices and private goods, the best decision rule may be individual choices (e.g., to take a trip or purchase a good), recognizing that incentive compatibility no longer directly applies. In these cases, survey designers should employ plausible decision-making frames and justify the use of the chosen decision format. Similarly, for public goods where a referendum vote may not be plausible, investigators should consider other plausible public decision-making processes that are binding and credible in the specific context.

5.5. Payment Vehicle

The payment vehicle, or the manner in which payments would be made, is a crucial component of SP questions. The literature has reached consensus on certain aspects and implications of payment vehicle selection. For example, binding (non-voluntary) payment vehicles are required for incentive compatibility and to prevent free-riding. However, there is no consensus on the selection of specific payment vehicles.

RECOMMENDATION 11: A payment vehicle should be selected to be realistic, credible, familiar and binding for all respondents to as great an extent as possible, and to ensure that payments are viewed as fixed and non-malleable. Payment vehicle selection should be informed by pretesting to minimize unintended effects on value estimates.

Several studies have investigated the effects of different payment vehicles on value estimates. The evidence is clear that the selection of a payment vehicle can influence these estimates.[48] Further, there is no single objective criterion that identifies what payment vehicle is best for a particular application. Valuing a public good using a referendum decision process may appeal to some type of national or local tax vehicle, but caution is warranted. For example, income and sales taxes can be problematic because income taxes may be paid by only a fraction of the public, and people can adjust how much they pay in a sales tax (or goods and services tax) by adjusting purchase decisions. A number of studies have used an (unavoidable) increase in the cost of living, which is realistic for many types of situations[49] and can deflect respondent focus on one specific type of payment. Voluntary and other non-binding payment mechanisms are not recommended due to a lack of incentive compatibility and the associated tendency of subjects to free-ride, although such mechanisms may be unavoidable in some contexts. Approaches such as provision point mechanisms have been shown to improve the demand-revealing properties of voluntary donations and other mechanisms that are not incentive compatible (Marwell and Ames 1980; Alston and Nowell 1996; Rondeau, Schulze, and Poe 1999; Poe et al. 2002; Rose et al. 2002).[50] Regardless of the payment vehicle used, investigators should document the chosen vehicle and the potential for the vehicle to result in under- or over-estimation of values.

The level of detail given to payment vehicle descriptions is also context-specific and

[48] For example, Johnston, Swallow, and Weaver (1999) find that respondents' trust in the payment vehicle affects marginal WTP and rates of substitution between non-monetary outcomes. Campos, Caparros, and Oviedo (2007) report that an increase in trip cost leads to higher recreation values than a site entrance fee. Morrison, Blamey, and Bennett (2000) find that the plausibility of a payment vehicle influences SP responses. Brouwer et al. (1999), in a meta-analysis of wetland valuation studies, find that income tax payment vehicles increase value estimates compared to alternative vehicles.

[49] The costs of many public policies are borne broadly through higher taxes, higher prices, lower wages and/or lower investment returns.

[50] A provision point is a decision rule such that if a certain number of people vote in favor of a program or sufficient funds are raised, the program will be implemented. All funds are returned if the program is not implemented.

should be chosen with input from pretesting. Greater detail can increase credibility, but at the potential cost of increasing the number of protests and/or respondents who view the payment as avoidable or malleable (Bateman et al. 2002, pp. 131-133). Other aspects of the payment vehicle, such as timing (e.g., single versus annual payments) and individual versus household payments should also be chosen based on the valuation context and pretesting. Although the frequency of payment has been shown to influence value estimates and there are compelling arguments for the use of periodic instead of lump sum payments (Egan et al. 2015), temporal aspects of realistic payment vehicles are context-dependent and must be considered. Lump sum payments contain implicit discounting assumptions by respondents while aggregation of payments over time requires investigator-imposed assumptions to be made on the discount rate.

Survey pretesting should not be used to dictate a payment vehicle based solely on respondents' preferences or desires. Such an action may lead to biased estimates of value. For example, respondents might prefer a non-binding vehicle that would enable them to avoid payment, such as a donation mechanism. In addition (or alternatively), a respondent's desired payment vehicle may not be realistic for the decision-making context. Such balancing of insights from pretesting and application-specific criteria applies to all elements of survey design.

5.6. Auxiliary or Supporting Questions

Auxiliary or supporting questions are often included in SP questionnaires to assist in understanding responses to value elicitation questions (Krupnick and Adamowicz 2006). These auxiliary questions can have multiple purposes, such as: (a) partitioning the flow of long sections of text; (b) helping to engage respondents as they process presented information; (c) evaluating whether (and how) respondents understand and/or accept information; (d) identifying protest

responses or other motivations for value elicitation responses; (e) providing information to evaluate validity; (f) evaluating respondents' perceptions of the survey instrument, (e.g., consequentiality, difficulty, neutrality, etc.); (g) understanding respondents' attitudes, opinions, behaviors, knowledge and experiences; and (h) identifying demographic, household or other characteristics. A subset of these may provide covariates used in valuation models to explain variation in responses to the value elicitation question(s).

RECOMMENDATION 12: SP questionnaires should include auxiliary questions to enhance the validity of the SP study and to evaluate the validity of responses to the value elicitation questions. These questions should have a specific purpose, established *ex ante*, and should be pretested to ensure that they serve the intended purposes.

Krupnick and Adamowicz (2006) review the types and uses of auxiliary questions. These questions are commonly used to support and enable evaluation of content and construct validity (Carmines and Zeller 1979).[51] Guidance on the design of questions to evaluate or enhance the content validity of an SP questionnaire can be obtained from the survey design literature (e.g., Marsden and Wright 2010; Groves et al. 2009; Dillman et al. 2014), as well as SP primers (e.g., Mitchell and Carson 1989; Bateman et al. 2002; Champ et al. 2017). These questions might consider such issues as whether the payment vehicle or other aspects of the survey were believed and accepted by respondents (Krupnick and Adamowicz 2006; Ivehammar 2009). Related uses include the identification of protests or other responses inconsistent with welfare estimation (Lancsar and Louviere 2006; Meyerhoff and Liebe 2008; Atkinson et al. 2012; Brouwer et al.

[51] As discussed below, content validity involves the procedures used to implement a study. "Content validity judgments encompass the entirety of the study [including] the clarity, interpretability and plausibility of the questions posed" (Bateman et al. 2002, p. 305). Construct validity, in contrast, considers how study results compare to hypothesis tests based on prior expectations (Bateman et al. 2002; Mitchell and Carson 1989).

2012). Auxiliary questions are also crucial for evaluating concerns identified during pretesting that cannot be fully resolved for all respondents during questionnaire design (e.g., regarding the believability, adjustment and/or rejection of the valuation scenario; Cameron et al. 2011).

Following guidance provided by Krupnick and Adamowicz (2006), each auxiliary question should be designed and pretested for a specific and identified purpose. It is important to consider the type of information provided by each auxiliary question and how it relates to value elicitation, considering both theoretical and statistical concerns. For example, responses to auxiliary questions may be endogenous to valuation responses and while a useful part of the survey, may have limited use in the estimation of valuation response equations. Also, even subtle differences in questions asked prior to valuation tasks may affect respondents' subsequent choices, as shown by Cai et al. (2010). The potential for endogeneity and confounding effects of auxiliary questions should be carefully addressed in econometric estimation. Finally, concerns of unknown or complex incentive properties also apply to auxiliary questions. It should not necessarily be assumed that these questions will always promote truthful responses.

5.7. **Ex-Ante Procedures to Enhance Validity**

There are multiple *ex ante* procedures that have been suggested to enhance the validity of SP value estimates (Loomis 2014). Commonly cited procedures include the use of cheap talk scripts, provision-point mechanisms, honesty oaths, visible choice sets, and information to enhance consequentiality and incentive compatibility. These are typically applied in an effort to reduce presumed over-estimation of values (List and Gallet 2001; Murphy, Allen, and Weatherhead

2005b), but may also avoid underestimation (Carson et al. 1996).[52]

RECOMMENDATION 13: SP valuation scenarios and valuation questions should be designed to enhance incentive compatibility and to encourage truthful responses. Approaches that enhance valuation scenario consequentiality are recommended.

As emphasized above, SP questionnaire designs should provide clear choices that are consequential and incentive compatible (Hurwicz 1986; Groves, Radner, and Reiter 1987; Varian 1992; Carson and Groves 2007; Vossler et al. 2012; Carson et al. 2014). Vossler et al. (2012) and Carson et al. (2014) find that consequential SP choices encourage truthful preference revelation. Incentive compatibility depends in part on whether the elicitation is viewed as consequential, including payment and policy consequentiality (Herriges et al. 2010). Following general survey practices, promoting consequentiality includes the provision of information on funding agencies and how the results of the survey will be disseminated. As discussed by Carson and Groves (2007), consequentiality is also affected by other aspects of survey design such as the plausibility of prices or goods, and details about how goods would be provided.

Cheap talk is an approach that evolved from experimental economics (Farrell and Gibbons 1989) to CV and CE applications (Cummings and Taylor 1999; Lusk 2003). The central tenet is a reminder of the hypothetical (i.e., likely non-consequential) nature of scenarios and the tendency of respondents to inflate value estimates. The incentive properties of cheap talk are not clear, however, and cheap talk does not always reduce value estimates (Murphy et al. 2005b; Aadland and Caplan 2006; Loomis 2014) suggesting that this method should not be applied

[52] *Ex-post* validity adjustments are discussed under the discussion of SP data analysis below.

without consideration of the implications for framing and consequentiality. In general, treatments that explicitly reduce consequentiality are not recommended.[53]

Other approaches that have been suggested include honesty-based methods, for example in which respondents are asked to sign a truth-telling oath (Jacquemet et al. 2013, 2016; Stevens, Tabatabaei, and Lass 2013). Those who do not agree to sign the oath may be excluded from the analysis, although this does not always influence results; the primary function of the oath script is to increase respondents' commitment and attention (Carlsson et al. 2013). However, the channels through which oaths affect behavior are unclear, and oaths have the potential to influence stated preferences in unintended ways.

The evidence is mixed (or limited) on these and other *ex ante* approaches to validity enhancement, and careful consideration should be given to utilizing any single approach in a survey. Moreover, we are aware of no research that considers the impact on valuation estimates when two or more *ex ante* validity approaches are included in a survey. Given equivocal evidence supporting any one (or combination) of these methods, we believe that the most promising *ex ante* approach remains a consequential design with a binding payment. We encourage research into ways that various approaches can be used to enhance validity, incentive compatibility and truthful preference revelation.

6.0 DATA ANALYSIS

Advancements in econometric analysis are a major theme in the SP literature. There are many parametric, semi-parametric and non-parametric modeling alternatives (Haab and McConnell

[53] Cheap talk must be recognized as a strategy that focuses respondent attention disproportionately on the cost of each alternative in the choice set(s). In general, one should proceed carefully when directing attention disproportionately to one or another attribute in any choice context (Cameron and DeShazo 2010).

2002; Train 2009; Scarpa and Alberini 2005; Hess and Daly 2014). In terms of estimation, SP data are no different from other types of economic data; the most appropriate estimator should be selected for the format of the data and the question(s) to be answered. Best practices for the estimation and interpretation of SP models parallel those for other types of statistical models with similar data structures. The following recommendations focus on issues that are specific to SP data analysis or for which guidance requires application-specific considerations.

6.1. Choice of Econometric Estimator

Multiple types of SP data are possible, each with analytical challenges that require specific econometric methods. Whereas econometric methods for analyses of some types of SP data have undergone little change in recent years (e.g., open-ended CV data), others have been subject to rapid development (e.g., discrete-choice data). Current modeling often employs simulation-based approaches to address specific types of response patterns (Lewbel 2000; Swait and Adamowicz 2001a,b; Scarpa, Thiene, and Marangon 2008a, 2009; Train 2009; Scarpa and Alberini 2005; Watanabe 2010; Lewbel, McFadden and Linton 2011; Hess and Daly 2013, 2014; Boeri, Scarpa, and Chorus 2014). For example, estimation often allows for preference (and/or scale) heterogeneity and correlated responses with various types of conditional, mixed or generalized logit models (Revelt and Train 1998; Train 2009; Fiebig et al. 2010; Scarpa and Alberini 2005).

RECOMMENDATION 14: No one particular model or set of models is recommended for all SP modeling. Econometric estimator selection should reflect the unique aspects of the data to be analyzed, the hypotheses to be investigated and how the estimation results will be used to support decision-making. Tradeoffs considered in estimator selection should be explicitly documented. Modeling should be informed by the utility-theoretic structure assumed to motivate

behavior within particular SP contexts. Utility-theoretic, behavioral, statistical, and other assumptions underlying model selection and specification should be made explicit.

Fundamental to SP analysis is a utility-theoretic foundation that guides the value to be estimated and the hypotheses to be tested. All SP analyses require assumptions regarding the structure of preferences. Basic axioms of choice suggest that (a) as the bid amount increases the proportion of people who would pay the amount should not increase and (b) WTP should be non-decreasing in an increase in quantity or quality, unless the individual is satiated and further increases provide disutility.[54] Budget constraints should also be relevant, acknowledging that estimated values are often a small percentage of income and, compared to gross income, an individual's operational budget may be lower (e.g., post-tax or discretionary income) or higher (e.g., due to wealth) and not observed. The availability or prices of substitutes or complements are also expected to influence values, but individually relevant substitutes or complements may be unclear. Beyond these basic considerations, additional structure on how preferences are framed is a matter of assumptions and context, and will vary across studies.

Survey design and initial data analysis should be preceded by a consideration of the functional specifications to be estimated, any uncertainties regarding model structure and the primary hypotheses to be tested.[55] If there is uncertainty regarding model specifications, it can be investigated through robustness checks during data analysis. It is also possible to apply formal approaches such as Bayesian model search and averaging or frequentist analogs (Layton and Lee

[54] This presumes that the individual cannot freely dispose of or discard the unwanted units of quality or quantity, as is the case with some types of public goods (e.g., large populations of wildlife causing a public nuisance).
[55] Some have advocated that analysts should pre-specify and publicly archive a detailed analysis plan before seeing the data (as in Finkelstein et al. 2012). Such practices can identify specifications and estimates that conform to *ex ante* research plans and those discerned only after exploration of the data. Results of both types can be valid and publishable, but the former are less subject to possible data mining and specification searching, leading to risks of incorrect statistical inference (Leamer 1983; Lovell 1983; Veall 1992; Layton and Lee 2006).

2006; Balcombe, Chalak, and Fraser 2009). This recommendation does not preclude analysts from exploring model specifications that were not envisioned *ex ante* (or from having data inform theory rather than vice versa); this can be a central part of data analysis. Rather, it is intended to ensure transparency in model development and the resulting statistical inferences.

Specifics of model structure will depend on the type of responses being modeled and other aspects of the data. For example, econometric analysis of open-ended CV responses usually involves the estimation of tobit models or other specifications with a spike in the distribution that allows for a corner-solution or censored outcome interpretation of zero-value ($0) responses (Kriström 1997; Haab and McConnell 2002). In contrast, contemporary SP questions typically yield categorical responses, leading to econometric models for limited dependent variables grounded in random utility theory (Manski 1977; Hanemann 1984; Maddala 1986; McConnell 1990; Adamowicz et al. 1998).[56] The core model for these applications remains the multinomial logit (MNL; also called conditional logit) model and its variants, although similar models such as multinomial probit may also be applied.

There are several common parameterizations of logit models. The use of mixing distributions can minimize the consequences of a failure of preferences to satisfy the independence of irrelevant alternatives (IIA) assumption in basic logit models. These generalizations can approximate any preference structure (McFadden and Train 2000). Identifying a correct specification for mixed logit models poses challenges, however, including the choice of appropriate mixing distributions. This choice can affect welfare estimates (Sillano and de Dios Ortúzar 2005; Meijer and Rouwendal 2006). Theory and intuition can sometimes guide the choice of distributions (Hensher and Greene 2003; Train 1998; 2009) and empirical

[56] Payment card (interval) data are now less common, and quantitative methods for these types of data have seen little change in the last few decades. Haab and McConnell (2002) summarize methods of analysis for those formats.

guidance on distributions can be gleaned by examining the posterior distributions of individual-specific coefficients (Huber and Train 2001; Greene, Hensher, and Rose 2005; Scarpa and Thiene 2005). Analyst judgement is required, however, leading to concerns over the robustness/fragility of estimates (Sillano and de Dios Ortúzar 2005; Layton and Lee 2006; Meijer and Rouwendal 2006; Balcombe et al. 2009; Johnston and Bergstrom 2011; Torres, Hanley, and Riera 2011).

Future econometric advances will surely provide modeling options unavailable today. Thus, we emphasize the importance of selecting estimating specifications that are consistent with a clearly specified utility foundation (necessary for welfare analysis) and that are capable of accommodating the unique aspects of the response data.

6.2. Modeling Heterogeneity

Preference heterogeneity has long been recognized in SP studies, from ad hoc inclusion of socio-demographic characteristics in model specifications to recent analyses that use sophisticated econometric approaches to model heterogeneity in preferences associated with both observable and unobservable factors (Train 2009; Scarpa and Alberini 2005). Although the relevance of preference heterogeneity varies according to the goals of the analysis, accommodating it during data analysis has multiple advantages. For example, unobserved heterogeneity can bias coefficient estimates based on MNL models under certain conditions (Train 1998; Van den Berg, Kroes, and Verhoef 2010).[57] Knowledge of preference heterogeneity can also help analysts understand differences in preference coefficients and identify whether sample selection biases affect aggregate estimates of benefits or costs, and can help decision-makers understand how

[57] Failure to account for scale heterogeneity (i.e., differences in error variance) can also distort results of various validity tests and other statistical comparisons (e.g., Haab, Huang, and Whitehead 1999).

benefits and costs are distributed across the population.

RECOMMENDATION 15: Analysis of SP data should allow for both observed and unobserved preference heterogeneity, and should consider the relevance of this heterogeneity for the use of study results to support decision-making.

This recommendation does not imply that all analyses of SP data need to model heterogeneity explicitly. However, analysts should consider whether and how heterogeneity may be relevant to consistent estimates of preference parameters, interpretation of estimation results, and the use of point estimates to compute aggregate welfare measures under historical or new conditions.[58] Given the wide array of econometric approaches that can accommodate preference heterogeneity, we do not recommend any particular model for all applications. For example, there may be discrete groups of people where each group shares a different set of preferences (suggesting a model with a different set of preference parameters for each group, or perhaps a latent class approach). Or, preferences may vary continuously across all individuals (suggesting a model with utility parameters that vary systematically as a function of observable respondent attributes[59] or a mixed logit model with randomly varying parameters[60]). Alternatively, there may be one shared set of preferences across all respondents, but the error term may have differing dispersions (suggesting an error component model).

Guidance concerning the types of heterogeneity that should be accommodated can be

[58] For example, mixed logit models often generate different estimates of mean WTP than do MNL models (van den Berg et al. 2010).
[59] Observable attributes frequently associated with systematic variation in utility parameters include demographic attributes such as income and education, along with spatial (or locational) attributes such as distance to affected areas or residence in a particular region (e.g., Sutherland and Walsh 1985; Pate and Loomis 1997; Hanley et al. 2003; Bateman et al. 2006a; Brouwer, Martin-Ortega, and Berbel 2010b; Campbell, Hutchinson, and Scarpa 2009; Rolfe and Windle 2012; Schaafsma et al. 2012; Jørgensen et al. 2013; Johnston et al. 2015).
[60] Models have been proposed in which both discrete and continuous distributions are present (Bujosa, Riera, and Hicks 2010).

obtained through consideration of the decision or change to be addressed, insights from previous research and focus groups, survey pretesting and other preliminary data analyses, and robustness checks during model estimation. Preference variations can also be confounded with variations in scale of the utility parameters (equivalent to variation in the scale of the error variance), implying the importance of estimators that allow scale to vary across respondents (Swait and Louviere 1993; Fiebig et al. 2010). The choice among estimators requires assumptions consistent with the intended use of model estimates and an understanding of the data in question.

Given the ease of estimating models that include preference heterogeneity, and risks associated with models that impose homogeneous preferences when this assumption does not hold, modeling preference heterogeneity should be standard practice. At the same time, model selection and estimation require investigator choices that should be justified and documented. In addition, while modeling preference heterogeneity may be consistent with the distribution of preferences in the population, it is important to recognize the limitations of some approaches in terms of welfare estimation (Train and Weeks 2005; Hole 2007; Daly, Hess, and Train 2012).

6.3. Balancing Model Parsimony and Complexity

SP data estimation involves tradeoffs between the use of parsimonious and more-complex models. We define a “parsimonious” model as one that requires fewer components and/or investigator-imposed assumptions or modeling decisions. We define a “complex” model as one that applies a larger number of these features, typically in an attempt to glean greater insight from the data. At the simplest level, a parsimonious model might involve nonparametric estimation of a model requiring few analyst assumptions while enabling investigation of the basic axioms of choice. Complexity is added by moving to a parametric specification, imposing a

nonlinear response function, adding covariates, etc.

RECOMMENDATION 16: Comprehensive SP data analysis should include both (a) the simplest, most parsimonious specifications with maintained hypotheses consistent with the basic axioms of choice and properties of the data, and (b) more-complex models that impose additional investigator assumptions on the structure of responses.

Manski (2008) proposes a bottom-up approach wherein the simplest model (i.e., one that minimizes investigator-imposed assumptions and decisions) is estimated first, followed by more complex models. This strategy is not necessarily straightforward for some types of SP analysis, because some of the simplest parametric models used for SP analysis (e.g., MNL) involve strong and often-violated assumptions (e.g., IIA). For this reason, we do not recommend a particular order for the estimation of parsimonious versus more complex models. We do, however, recommend that both should be used as part of comprehensive data analysis.

Nonparametric models are an important part of this strategy (Haab and McConnell 2002; Watanabe and Asano 2009; Watanabe 2010; Kaul et al. 2013). These would be accompanied by simple parametric models, together with more flexible alternatives that may be more sensitive to specification choices, or in which unique identification of effects may be difficult such as generalized multinomial logit models (Fiebig et al. 2010; Hess and Rose 2012). Although the assumptions implied by simpler econometric models—such as the IIA assumption within MNL models—may be rejected, such models can provide a useful baseline against which the results of more-sophisticated models can be compared.

Analysis should also include assessment of the primary design variables and their impact on estimation outcomes, apart from effects of other covariates. In discrete-choice CV, this

involves only the bid variable. In CEs, all design attributes should be included, as the random assignment of choice sets to respondents typically guarantees exogeneity of the mixes of these attributes as determinants of individual choices (Carson and Hanemann 2005).[61] In repeated choice contexts, where choice sets are randomized both within and across respondents, it can be desirable to report estimates using only the first choice by each respondent.[62] This choice will not be affected by any previous choices and a first-choice-only analysis will reduce the potential influence of strategic behavior associated with respondent-inferred non-independence of choice cases (Day et al. 2012; Vossler et al. 2012). However, if learning is involved and desirable, responses to the first question may not provide the best coefficient estimates. In such cases, subsequent models using all of the data may then be used to evaluate the role of potentially relevant covariates that may ultimately affect value estimates.

Many published SP studies facilitate estimation by assuming a utility function that is linear and additively separable (with constant marginal utilities). Although such functions may serve as a useful local first approximation, these implicit assumptions will not always hold. Among the concerns in this area is the likelihood that preferences will exhibit non-linearity (e.g., diminishing marginal utility or non-constant marginal rates of substitution between attributes). Such possibilities can be accommodated using richer specifications for preference or welfare functions. More complex models may be necessary to accommodate preference and/or scale heterogeneity, and also to accommodate more complex preference functions at the individual level. In addition to general concerns about functional form and its correspondence to welfare theory, other aspects of model specification should be considered—based on the weight of the

[61] If the design makes all attribute levels both exogenous and independent from each other, there will be no omitted variables bias if subsets of attributes are excluded from the model.
[62] Such analysis must be facilitated by an experimental design that supports analysis of first responses in isolation.

evidence from prior work and insights from pretesting.[63, 64]

It is unnecessary and may not be possible to report the results of all estimated models within published articles, although it is often useful to report the results of multiple model specifications. However, results should be retained to inform future modeling and as evidence of due diligence in data analysis, and can be documented in footnotes. Many journals also accept "supplementary materials" for published papers, which are archived online for interested readers.

6.4. Behavioral Response Anomalies

Like all human behavior, responses to SP questionnaires may be subject to various types of behavioral or response "anomalies." Such anomalies have received considerable attention in the literature (see, e.g., Carson 2012; Hausman 2012; McFadden 2014). One must be careful when discussing behavioral anomalies in SP responses. Are the observed patterns (a) "anomalies" in the sense that respondents are not revealing their true preferences, or (b) "anomalies" in the sense that response patterns are simply inconsistent with the structure and assumptions imposed by the investigator? It is patterns of type (a) that are of most concern here. These anomalies can be overt (e.g., different types of strategic or protest behavior) or inadvertent (e.g., anchoring on bid amounts, or on the first question in a sequence of valuation questions). Further, many anomalies may be non-systematic and individual-specific, and thus may simply contribute to (essentially) random noise in response data. However, more severe and systematic anomalies that arise across a large portion of the sample can have a non-trivial influence on model estimates.

[63] For example, one should test for the relevance of alternative-specific constants (ASCs), where applicable, and particularly in CEs with labeled or status quo alternatives; failure to do so can bias parameter estimates (Adamowicz et al. 1998; Kerr and Sharpe 2006).

[64] The potential for non-linearity of preferences, including attribute interactions as well as preference heterogeneity, should be considered in survey development (including sample and experimental designs) so that these considerations can be adequately assessed in the estimation phase.

RECOMMENDATION 17: When prior research or pretesting indicates that undesirable response anomalies may be influential, data analysis should investigate these anomalies to determine whether they significantly affect SP responses. For example, analyses should consider whether protest or outlier responses are influential and conduct follow-up analyses, as warranted, to identify their effects. Where evidence suggests that information in scenarios may conflict with prior beliefs, data analysis should investigate the possibility that respondents may not interpret scenarios exactly as intended. However, every analysis need not evaluate all possible anomalies. Consideration should be given to whether response anomalies merely add random noise to value estimates or whether they have systematic effects that can distort these estimates.

Where clear symptoms of systematic anomalous responses have been identified from similar studies in the literature or during pretesting, SP studies should be designed to avoid these anomalies. When this is not possible, the surveys should be designed to investigate anomalous responses and analyses should use the information to investigate the effects. Many of the anomalies discussed in the literature fall under the general classification of behaviors that may not be fully consistent with simple, linear, fully compensatory, discounted expected utility maximization. These often arise from individuals applying simplified decision rules to reduce the cognitive burden presented by a survey (Kahneman and Tversky 1979; Schoemaker 1982; Mazzotta and Opaluch 1995; DeShazo and Fermo 2002). Respondents' lack of experience with the type of choice being proposed may lead to inadvertent, experimentally induced errors in responses to choice questions (Boyle et al. 1997; Bateman et al. 2001; Bateman et al. 2004; Day et al. 2012, Czajkowski, Hanley, and Nyborg 2015). These errors may decrease with a respondent's real-world experience with similar goods, or over the course of a sequence of

questions (Boyle et al. 1993; Braga and Starmer 2005; Bateman et al. 2008a; Brouwer et al. 2010a; Day et al. 2012).

Another type of anomalous response is scenario adjustment or rejection, in which respondents do not interpret scenarios as intended and thus value something different from the intended item or outcome (Carson et al. 1994; Cameron 2005a,b; Burghart, Cameron, and Gerdes 2007; Cameron et al. 2011; Cameron and DeShazo 2013). Rather than altering the scenario, some respondents may also universally choose either the status quo or a non-status quo alternative; this may be interpreted as another form of scenario rejection (Kristrom 1997; Carson 2000, von Haefen, Massey, and Adamowicz 2005). The primary means to reduce scenario rejection and adjustment, or non-participation in the choice, is through adjustments in survey design. However, if pretests show such issues are unavoidable with some respondents, the questionnaire should include debriefing questions to allow the effect on value estimates to be evaluated during data analysis, and ameliorated if possible (Adamowicz et al. 1997; Cameron 2005a,b; Burghart et al. 2007; Cameron et al. 2011; Kataria et al. 2012; Cameron and DeShazo 2013). The causes, consequences and mitigation of scenario adjustment and rejection are areas for which additional research is warranted.

Response anomalies may also be associated with experimental or survey design features. Examples include bid anchoring (Boyle et al. 1997; Bateman et al. 2001), insufficient responsiveness to bid amounts or scope (Diamond and Hausman 1994; Holmes and Kramer 1995; Blamey et al. 1999; Powe and Bateman 2004; Veisten et al. 2004; Heberlein et al. 2005), sequencing effects (Bateman et al. 2004; Day and Prades 2010; Day et al. 2012) and attribute non-attendance (Scarpa et al. 2009, 2010; Cameron and DeShazo 2010; Campbell et al. 2008a, 2011), among others. Concerns such as these should be addressed during survey design and

pretesting. If evidence suggests that these effects persist and are influential (despite efforts to minimize them during survey design), they should be investigated as part of data analysis.

Outliers and protest responses can be examples of anomalous responses, or they can emerge for different reasons (Halstead, Luloff, and Stevens 1992; Jorgensen et al. 1999; Jorgensen and Syme 2000; Bateman et al. 2002; Meyerhoff and Liebe 2008). These types of responses should be considered in pretesting and subsequently during data analysis. However, the evidence is mixed concerning methods to identify and account for protest responses. Common identification approaches include the use of open- or closed-ended debriefing questions (Bateman et al. 2002; Meyerhoff, Bartczak, and Liebe 2012), statistical outlier detection (Belsley et al. 2005), and the identification of respondents who answer either always "yes" or always "no" to every valuation question (Kristrom 1997; Carson 2000; von Haefen et al. 2005; Boxall et al. 2009; Meyerhoff and Liebe 2009).[65] Heckman and other types of sample selection models have been proposed as a means to address sample selection due to protest bids or the removal of apparent protest bids by analysts (Edwards and Anderson 1987; Messonnier et al. 2000; Brouwer and Martin-Ortega 2012). Identification of protest responses and data outliers often requires the analyst to use subjective judgment—there are frequently no clear-cut decision rules or criteria for such identifications (Jorgensen and Syme 2000; Bateman et al. 2002; Meyerhoff and Liebe 2006; Rollins et al. 2010). The main concern for SP data analysis is how protest responses and outliers should be handled in the estimation of values. Approaches include dropping observations, conducting analyses with and without these observations, and developing models that attempt to control for factors that affect protests (Meyerhoff et al. 2012, 2015).

Although the literature is clear that protest responses of various types may be a concern,

[65] As described above, parallel behavior in CEs occurs when respondents either choose the status quo (or reject the status quo) on all choice occasions, regardless of attribute levels.

there is no agreement on a single set of best practices to address these problems. Transparency in the identification and treatment of suspected protests is therefore essential, as is sensitivity analysis. Questionnaires should include debriefing questions to identify possible protest responses, recognizing that these questions do not always uniquely identify protesters. If analyses of responses to these follow-up questions (or other aspects of response behavior) suggest that protest responses may be pervasive and influential, sensitivity analyses are warranted. Likewise, if there are substantial numbers of apparent outliers among the responses, the approach should parallel the treatment of outliers in any type of econometric analysis.

Behavioral anomalies are not unique to SP studies; they also occur in RP settings (Bateman et al. 1997a,b, 2006b; Braga and Starmer 2005; Bateman, Munro, and Poe 2008b; Carson 2012). Anomalies such as these are the essence of behavioral economics and an important determinant of how marketing affects consumer choices (DiClemente et al. 2003; Kahneman 2003; Ho, Lim, and Camerer 2006). Thus, while a goal of SP design and analysis is to minimize the effects of anomalies, it is important to remember that there is likely no behavior, even in markets, that is universally consistent with the simple textbook neoclassical microeconomics paradigm. It is also possible that a respondent might answer one or more questions in manner that appears anomalous, but answers other questions in a way that is consistent with rational economic choice. How such information should be used to evaluate credibility has not been sufficiently addressed in the literature.

6.5. Value Estimation

Efficient and unbiased value estimation is among the primary goals of most SP analysis. Achievement of this goal can be jeopardized by unsuitable statistical models, inappropriate

functional specifications, or incorrect investigator assumptions. Moreover, efficient and unbiased parameter estimates do not guarantee that welfare estimates derived from the estimated parameters have similar properties.

RECOMMENDATION 18: Reported welfare estimates should, at a minimum, include estimates of central tendency and dispersion. Methods used to calculate welfare measures should be transparent and should ensure that estimates are theoretically and statistically well defined. This applies to all reported moments, quantiles, etc.

Computation of value estimates is not always straightforward. Consider, for example, the case of welfare estimates derived from a mixed logit model. These estimates depend on the imposed distributional assumptions for each of the coefficients. Conventional preference-space utility specifications often imply implausible distributions of welfare estimates (in many cases with undefined moments), given that the typical estimate of WTP is calculated as the ratio of two (perhaps jointly) distributed coefficient estimates (Train and Weeks 2005; Scarpa, Thiene, and Train 2008b; Daly et al. 2012).[66] Attempts to address this problem via strong and generally unrealistic assumptions (e.g., imposing a fixed coefficient on the cost attribute for all individuals and groups) cannot be universally recommended for general application, although assumptions such as these may be appropriate in some cases.

One increasingly used alternative is the estimation of indirect utility in WTP-space, in which the distribution of the welfare measure is modeled directly (Cameron and James 1987; Train and Weeks 2005; Scarpa et al. 2008b). The fit of such models is not always as good as that

[66] Similar problems can occur even in the earlier class of multinomial or other non-simulation based logit models if the cost coefficient is not significant (Hole 2007).

achieved with preference-space analogs, and estimation can be challenging.[67] Nonetheless, analysts should consider sensitivity analysis of WTP distributions to the two approaches when preference-space WTP distributions raise concerns or when implied WTP distributions do not have finite moments. The preferred model for computing value estimates should be identified and reasons for selection documented. Regardless of the approach, it is incumbent upon analysts to clarify their assumptions and the implications of the approaches they use, to ensure that the reported welfare measures are well defined.

6.6. Using Data from Auxiliary and Supporting Questions

In SP studies, responses to supporting, debriefing or follow-up questions have been used (a) as covariates within models, (b) to segment or restrict the sample, (c) to develop inferences about the validity of valuation responses, or (d) to support *ex post* validity adjustments to valuation responses. Responses used for such purposes include questions on attitudes, knowledge or experience; acceptance and understanding of scenarios; uses of the goods in question; certainty in responses (see review in Champ, Moore, and Bishop 2009), emotions (Araña, León, and Hanemann 2008; Araña and León 2008), and truth-telling oaths (Jacquemet et al. 2013, 2016), among others.

RECOMMENDATION 19: Responses to supporting or debriefing questions are important components in an SP study. The use of data from the supporting and debriefing questions should

[67] Unobserved heterogeneity in the marginal utility of money (i.e., the cost coefficient) can also be modeled by using a means of mixing distributions with a finite number of points (e.g., latent class models). Or, one may model marginal utility as a function of income categories, for example using a spline function as suggested in (Morey, Sharma, and Karlstrom 2003). Even in this case, very low cost coefficient values in the denominator of the WTP formula may cause high ratios and implausibly high marginal WTPs (Train and Weeks 2005; Daly et al. 2012). Furthermore, if the marginal utility of money is not constrained to be strictly positive, for example by estimating its logarithm, a tiny negative value for the denominator may produce a huge negative estimate of WTP.

be accompanied by clear theoretical, survey design or empirical arguments explaining and justifying their use. These data may also be endogenous when used as explanatory variables in model estimation. Analysis should proceed with consideration for potential endogeneity and related concerns such as measurement error.

Data from supporting questions can be useful to help analysts understand variation in values across respondents, to evaluate the validity of valuation responses, and, in some cases, to make *ex post* adjustments to valuation responses to enhance validity. However, when data from supporting questions are used as covariates in valuation models, consideration must be given to whether these variables are endogenous to valuation responses. Analyses that overlook this endogeneity risk the provision of biased or otherwise misleading value estimates. Endogeneity is a particular concern when the data are collected from questions asked after the valuation questions, as responses to these questions may be influenced by how subjects answered the valuation question(s). Endogeneity is less of (or not) a concern when questions elicit respondent characteristics that are clearly exogenous to the valuation response (e.g., demographics[68]), or when answers are used for sample segmentation or to produce descriptive statistics to enhance the credibility of welfare estimates.

Other types of applications can raise endogeneity concerns. These include approaches based on the use of auxiliary questions to screen or adjust the data in an attempt to reduce presumed biases. Variables constructed from auxiliary questions have also appeared as covariates in econometric models, for example as multiplicative interactions within an estimated utility function. Examples include variables measuring: (a) respondents' perceived certainty in

[68] Examples include variables such as age, gender and ethnicity, which are (for the most part) not a result of other choices made personally by the same respondent.

valuation responses, (b) the extent to which respondents view value elicitation question as consequential, (c) whether respondents are willing to take oath of honesty, and (d) respondents' environmental beliefs or attitudes (Champ et al. 1997; Champ and Bishop 2001; Blumenschein et al. 2008; Blomquist, Blumenschein, and Johannesson 2009; Ready, Champ, and Lawton 2010; Herriges et al. 2010; Jacquemet et al. 2013, 2016). Among the potential causes of endogeneity in such cases is the fact that responses to value elicitation questions are motivated to some extent by factors unobserved by the analyst, and responses to auxiliary questions such as these may be motivated by the same factors. A few studies have highlighted and accommodated the resulting endogeneity concerns (Cameron and Englin 1997; Herriges et al. 2010).[69] However, the issue is often overlooked in SP data analysis.

Variables created from responses to auxiliary questions may also be subject to measurement error, as when underlying latent attitudes are measured with error using Likert-scale questions and subsequent factor analysis (Train 1987). The combination of endogeneity and potential measurement error in auxiliary data has received limited attention in the SP literature (Train, McFadden, and Goett 1987; Mariel, Meyerhoff, and Hess 2015; Hess and Beharry-Borg 2012). Recent advances in hybrid choice models such as the integrated choice and latent variable (ICLV) model are designed to address such concerns (e.g., Ben-Akiva et al. 1999; Hess and Beharry-Borg 2012; Czajkowski et al. 2015; Dekker et al. 2016).

Another challenge with data adjustments based on auxiliary questions (e.g., recoding, calibrating or excluding responses) is the absence of objective and theoretically defined criteria.

[69] While endogeneity has been considered in the labor economics literature (Angrist 2001; Carrasco 2012), it is often overlooked in SP analyses. If the attributes of each choice scenario are exogenously and randomly assigned, there can be no relationship between their levels and the characteristics of the respondents (assuming the absence of associated selection bias). This reduces some types of endogeneity concerns compared to those potentially encountered in models of observational RP data.

The standard for data re-coding, calibration or exclusion might vary across applications, and it is not possible to conduct a companion criterion-validity investigation for every study to identify the threshold for such manipulations. Moreover, there is no guarantee that the resulting subsamples of respondents will remain representative of the population of interest.

In summary, although the use of auxiliary data in SP studies is an important aspect of data analysis, consideration must be given to how these data are used within modeling. The use of auxiliary data is an area that begs for clear conceptual, theoretical and econometric foundations. These are important areas for future research.

6.7. Sample Representativeness and Value Aggregation

Much of the SP literature emphasizes results drawn from samples of convenience to investigate methodological or theoretical considerations. Unrepresentative samples of convenience can document the presence of preference or welfare patterns that are likely to be present, to at least some degree, in the general population. However, the resulting welfare measures have unknown generalizability (e.g., in terms of population-level mean or aggregate WTP) (Edwards and Anderson 1987; Whitehead 1991; Whitehead et al. 1993, 1994; Messonnier et al. 2000; Krupnick and Evans 2008). However, any SP estimate may be used in future decision-making applications via benefit transfer. Thus, knowledge of the sample frame and sampling strategy, as well as respondent characteristics, is necessary to use individual estimates of central tendency to compute aggregate estimates. Although it is not necessary for all publications of SP results to include formal assessments of representativeness, sufficient data should be available to enable at least minimal assessments, should the results be used to inform decisions.

RECOMMENDATION 20: The generalizability of value estimates from SP studies should be

documented. Analyses striving to produce decision- or policy-relevant estimates should include assessments to support the generalizability of value estimates to the sampled population. If studies do not seek to measure aggregate values, sufficient information should be provided to permit others to do so, or the analyst should explain why this would be unwise. If sample representativeness is unknown, the study should make this clear. In all studies, respondent characteristics should be documented in terms of standard socioeconomic characteristics as well as key application-specific characteristics. Calculation of welfare measures for policy guidance should recognize potential effects of sample selection, preference heterogeneity and the extent of the market. Any modifications in value estimates to address these considerations should be documented.

Given widespread concerns about low response rates, the survey literature increasingly recommends formal, *ex post* analyses of nonresponse bias (Groves 2006; National Research Council 2013). Nonetheless, such analyses are rarely conducted within the SP literature, in part due to a lack of data on non-respondents. However, some data on non-respondents (or the sample frame) can usually be obtained without costly follow-up efforts. For internet surveys, analysts often can obtain panel profile characteristics. Summary statistics on the observable demographics of respondents can also be compared with known population characteristics from sources such as national Census data. To enable such evaluations, survey designers should include questions in the survey that match the format used in the previous data collection effort for which characteristics are to be compared. Information documenting households or individuals who were invited to participate in the study should be preserved, including those who declined to participate or who withdrew (or were dropped due to missing data) prior to the final sample. When possible, data such as these can be supplemented with information from more involved

and costly follow-up contacts of non-respondents.

Comparisons such as these can identify differences between the population and sample in terms of observable characteristics, and can be used to re-weight response data to better represent the population (i.e., raking). However, these assessments provide little insight into the representativeness of a sample in terms of unobservable characteristics. To discern whether there is evidence of a correlation between unobservable characteristics and survey responses, a Heckman-type selection model can be estimated (Heckman 1979). There are various applications in the SP literature (e.g., Edwards and Anderson 1987; Messonnier et al. 2000; Brouwer and Martin-Ortega 2012). This approach can be applied to continuous and binary response data, but selection models have not been developed for more complex response data and panel data for limited dependent variables (Yuan, Boyle, and Wen 2015). For example, formal sample selection corrections for conditional logit models are not generally available.[70] In cases where formal methods do not yet exist, approximations may be used to gain insight into possible relationships between response propensities and estimated preferences (Cameron and DeShazo 2005, 2013; Yuan et al. 2015).[71]

In addition, respondents who completed the survey early in the field period (or with fewer contact attempts) can be compared with those who completed it later (or after more contact attempts) (e.g., Curtin et al. 2000; Johnston 2006). Some surveys also use alternative measures to

[70] If estimating a strictly reduced-form model, and a classical linear multiple regression model will suffice to reveal whether a potential relationship appears to be present in the data, it is possible to adapt this model to a Heckman selectivity specification. For choice scenarios, with multiple alternatives and multiple attributes, with preference parameters estimated by some variant of a structural conditional logit model, rigorous selection-correction algorithms are less readily available. Barrios (2004) offers "generalized sample selection bias correction under RUM," but the RUM part of the model concerns the selection equation, not the outcome equation.

[71] An ad hoc strategy for assessing the sensitivity of a choice model's parameter estimates to differing response propensities or probabilities requires that a relatively rich response/nonresponse model be estimated using the entire targeted (presumably) general-population sample. If there are statistically significant differences in response propensities, it can be informative to see whether a higher response propensity is associated with systematic differences in the preference parameters implied by the stated-preference exercise.

assess the likelihood of nonresponse bias, such as the R-indicators advocated by Schouten, Cobben, and Bethlehem (2009) and imbalance indicators proposed by Särndal (2011). These measures assess how well the respondents in a survey represent the population of interest.[72]

To enable analyses such as these, all SP studies, even if not designed to support decision-making, should include at least a minimal set of questions to collect sociodemographic information of the type commonly used to evaluate sample representativeness, plus any application-specific questions that would be relevant to such an evaluation. Summary statistics for all such data about households or individuals who participated in the study should be documented. If this information cannot be included in all published articles (e.g., due to binding page limits), it can be maintained in other readily accessible locations such as online appendices.

The use of WTP measures to guide decisions frequently requires aggregation to a population. Several approaches have been used in the literature to aggregate WTP. These can produce different estimates (Loomis 1996, 2000; Morrison 2000; Vajjhala et al. 2008). The most common approach is to scale sample-average WTP to the relevant population, adjusting for sample selection related to observable factors if necessary (e.g., using weights for various demographic and sampling features). The potential for spatial welfare heterogeneity should also be considered, as this can have significant and often underappreciated implications for welfare aggregation (Bateman et al. 2006a).[73] However, such approaches do not incorporate the potential for systematic, otherwise unobservable non-response bias associated with the study topic itself

[72] The R-indicator approach uses a response propensity model to estimate the probability that each sample case will respond to the survey; these typically are based on variables available on the sampling frame as well as data obtained in the course of carrying out the survey. The R-indicator is a function of the variability of the estimated propensities, $1 - 2 \times S(\hat{p})$, where $S(\hat{p})$ is the standard deviation of fitted propensities.

[73] Studies illustrating potentially relevant aspects of spatial welfare heterogeneity include Sutherland and Walsh (1985), Pate and Loomis (1997) Hanley et al. (2003), Bateman et al. (2006a), Brouwer et al. (2010b), Campbell et al. (2008b, 2009), Rolfe and Windle (2012), Schaafsma et al. (2012), Jørgensen et al. (2013), Johnston and Ramachandran (2014) and Johnston et al. (2015), among others.

(e.g., if people with higher values or more extreme views are more likely to respond).

An extreme alternative to sample weighting is to assume that all non-respondents place zero value on the change, resulting in a conservative estimate of aggregate value. A less conservative approach is to categorize non-respondents into those likely to have zero values and those likely to have values similar to respondents (Morrison 2000).

Given multiple possibilities for addressing otherwise uncorrected sample selection during benefit aggregation (where none of these methods are unambiguously supported by the literature), the most transparent approach is to aggregate benefits according to multiple assumptions as part of sensitivity analysis, while providing as much evidence as possible on the presence and extent of selection biases. This approach makes investigator aggregation assumptions explicit and provides insight concerning the robustness of estimates to the aggregation procedure. Of course, these actions are unnecessary if selection biases are not a concern. The best approach, and top priority, is to design a high-quality survey with pretesting and careful implementation to minimize the need for *ex post* adjustments in value aggregation.

7. VALIDITY ASSESSMENT

Much of the published research investigating SP methods has focused on the validity and reliability of value response data. Investigations have demonstrated that SP data are reliable (Loomis 1989; Teisl et al. 1995; Bliem, Getzner, and Rodiga-Laßnig 2012; Mørkbak et al. 2015). The primary issue of contention has been validity, and SP methods have been subjected to an extensive array of validity tests (Smith 2006b; Kling et al. 2012; de Bekker-Grob et al. 2012; Carson et al. 2014). The prevalence of SP validity studies suggests a consensus that such analyses are important.

As a precursor to our recommendations on the topic, this section provides a brief introduction to validity assessments within SP studies. Some individuals view hypothetical bias as the overarching validity issue facing SP methods; a long-standing concern has been that SP studies overestimate values (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005a).[74] The conclusion of hypothetical bias is largely based on studies in which an actual cash transaction (generally for a private good in an experimental setting) provides a benchmark against which SP estimates are compared. There is disagreement about the implications of these findings. Kling et al. (2012) and Carson et al. (2014) note that few of these studies satisfy incentive compatibility requirements for truthful preference revelation, and it is unclear how extent experimental payments compare to a "true" (or criterion) value.[75] Moreover, research using actual voting behavior as the criterion for comparison suggests that hypothetical bias is not universal (Vossler and Kerkvliet 2003; Johnston 2006; Vossler and Watson 2013). Assessing the validity of any SP study involves more than just a criterion validity test of "hypothetical bias."

There are three basic lines of validity investigations (Carmines and Zeller 1979; Mitchell and Carson 1989; Bateman et al. 2002; Smith 2006b; de Bekker-Grob et al. 2012; Bishop and Boyle 2017). *Content validity* reflects the appropriateness of the procedures used to design and implement a survey, the content of the survey instrument itself, data analysis procedures, and study reporting. It is generally evaluated using qualitative assessments of survey design and implementation with respect to the conditions to be valued, theoretical definition of the value to

[74] Much of this work has been conducted in the context of CV, but similar results have been found in CEs (Lusk and Schroeder 2004; Moser, Raffaelli, and Notaro 2013).

[75] Many criterion validity studies use an inconsequential experimental treatment to elicit behavior from which to compare cash transactions. There is mounting evidence that such a setting does not characterize the response incentives for many SP survey respondents, and validity tests conducted in more representative field environments have challenged the findings from standard experiments on hypothetical bias (Carson et al. 2014). Observed cash payments for public goods face additional challenges related to the potential for free-riding.

be estimated, the use of design and implementation features shown to be desirable in prior research, and pretesting of the survey instrument (Mitchell and Carson 1989; Bateman et al. 2002; Bishop and Boyle 2017). The content validity of any SP study depends on adherence to best-practice guidance concerning design and implementation, such as that proposed here.

Construct validity considers whether results from a nonmarket valuation study satisfy hypothesis tests based on prior expectations such as those informed by economic theory, previous empirical studies, issues identified in survey pretesting, and professional experience. Most validity tests in the literature are tests of construct validity. In addition to scope tests (tests of whether values are sensitive to changes in quantity or quality; Carson and Mitchell 1993; Carson 1997; Heberlein et al. 2005), for example, investigations have been conducted to examine the consistency of SP response data with theoretical considerations such as transitivity and other behavioral axioms (e.g., Johnson and Matthews 2001; Ryan and Bate 2001; Ryan and San Miguel 2003; Smith 2006b). Many different types of SP construct validity have been evaluated.

Convergent validity tests are a special case of *construct validity* and involve comparisons of SP estimates with parallel RP estimates, although it is acknowledged that neither may provide a criterion value (Mitchell and Carson 1989; Carson et al. 1996; Clarke 2002). Comparisons over different types of SP methods and elicitation formats have also been conducted (e.g., Boyle et al. 1996; Ready, Buzby, and Hu 1996; Hanley et al. 1998; Ryan and Watson 2009).

Criterion validity considers how SP value estimates compare to a presumed true value or criterion, and are the basis of many of the hypothetical bias assessments noted above. The standards for comparison may include experimental research (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005b) and voting comparison studies (Vossler et al. 2003; Vossler and Kerkvliet 2003; Johnston 2006; Vossler and Watson 2013). However, there are

disagreements over whether comparisons to some types of experimentally elicited values reflect criterion validity tests, or are instead a type of convergent validity test (Bateman et al. 2002, pp. 317-18; Carson et al. 2014; Ryan et al. forthcoming). Although tests of content and construct validity may be conducted using data from a single SP study, criterion validity tests require data from two or more parallel studies or data sources. Hence, although criterion validity tests are an important part of the literature, they are not feasible as part of most SP analyses. Moreover, if criterion values are available, there may be no need for SP analyses to inform decisions.[76]

7.1. Conducting and Interpreting Validity Tests

Validity is a key element of any empirical method, and all SP studies should demonstrate validity across multiple dimensions. However, validity tests must be interpreted *in context*, and with regard to issues such as (but not limited to) incentive compatibility, the effects of study design features, and differences between public and private good applications.

RECOMMENDATION 21: Analysis of SP data should include a set of core internal validity assessments, including formal tests of construct validity and evaluations of content validity. Validity test results should be interpreted within proper theoretical and empirical contexts, including recognition of potential confounding influences, the effects of study design, and the role of investigator-imposed assumptions.

This recommendation may be summarized as follows—all SP studies should include a minimum set of properly conducted, interpreted and reported validity tests. Studies should document study design and implementation procedures, along with the features used to enhance

[76] The exception may be when SP methods are necessary to forecast value estimates under new conditions that are well beyond the range of conditions observable in the available RP data.

the content validity of value estimates. Each study should also include a carefully considered set of construct validity investigations to support the validity of estimates, as appropriate to the context. These investigations could be based on issues of concern documented in previous studies (either relevant to all SP studies or to the particular type of application) and identified in survey pretesting.

Many different internal tests of construct validity are possible, but it is not practical or even relevant for any one study to conduct all possible investigations. Test results must also be interpreted in the appropriate context (Smith 2006b). For example, internal and/or external scope tests can be an important part of validity testing, but must allow for valid circumstances in which value estimates may or may not demonstrate responsiveness to scope (Rollins and Lyke 1998; Heberlein et al. 2005). The implementation and interpretation of scope tests has been an issue of debate for more than two decades, and economists looking at the same evidence have come to opposite conclusions with respect to validity (Hausman 2012; Kling et al. 2012).

Other analyses of survey response data might evaluate respondent perceptions of valuation questions, for example whether respondents perceived the valuation response to be consequential, payments to be binding, and scenarios to be plausible. Additional examples of construct validity tests include analyses of attribute nonattendance for CEs, protest responses, and serial non-participation (see citations to these topics above), among many others. Criterion validity testing is not feasible for each and every SP study. All the same, criterion validity assessment, where possible, should be an important component of future research.

When interpreting validity tests (including all convergent and criterion validity tests), it is important to recognize that all decision-making is contextual. Different types of evaluations introduce different contexts and potential confounds, and possible effects on validity test results

should be identified. For example, in the context of convergent validity tests, it should be recognized that RP methods are simply value estimation alternatives and the resulting estimates are not necessarily superior to SP estimates (Randall 1994).

Some have also recommended the use of more rigorous evaluations of construct validity such as the adding-up test (Diamond 1996; Diamond and Hausman 1994; Desvousges, Mathews, and Train 2015) and sufficient responsiveness to scope (Desvousges, Mathews, and Train 2012). These tests require investigator-imposed assumptions on preferences that go beyond the basic axioms of choice.[77] Testing assumptions such as these can be informative, but failure to meet these assumptions does not necessarily provide evidence against SP validity in general. Tests such as these are routinely violated in market settings (e.g., Bateman et al. 1997a,b), but these findings do not necessarily invalidate the use of RP data for welfare analysis. Partly as a result of these concerns, there is no consensus as to whether specific construct validity tests should be applied broadly as a means to validate or invalidate any given value estimate (Arrow et al. 1993; Hanemann 1994; Smith and Osborne 1996; Carson 2012; Haab et al. 2013). If adding-up, (sufficient) scope sensitivity, and similar tests are demonstrably feasible and relevant, then they can be applied along with other methods to evaluate responses, with the role of investigator-imposed assumptions made clear.

Underlying the challenge for SP validity testing is the lack of general agreement on whether results from individual studies (or sets of studies) should be interpreted as evidence for or against the validity of the method in general. Recognizing this lack of agreement over what constitutes an acceptable validity test for SP studies, we recommend continued investigation of

[77] In stylized terms, the adding-up test may be summarized as such: For three increasing levels of a single good A (A_1, A_2 and A_3), is an individual's WTP to move from A_1 to A_3 equal to their WTP to move from A_1 to A_2, plus their WTP to move from A_2 to A_3.

both current and new tests as an important area for future research.

7.2. Weight of Evidence in Validity Testing

Validity is a complex topic, and consideration of the weight of the evidence is crucial to make an educated decision on validity (Smith 2006b; Bishop and Boyle 2017). Beyond the contextual nature of validity tests discussed above, no single type of validity assessment provides a complete perspective, and it is common to use various tests to evaluate content, construct, and sometimes criterion validity (Mitchell and Carson 1989, pp. 190-193). Validity should be considered within the context of procedures used in within each study (e.g., to enhance content and construct validity), the findings of prior research, and the valuation topic being investigated.

RECOMMENDATION 22: Assessment of the validity of any study or valuation method should consider the weight of the available evidence and should not depend on the outcome of a single test or investigation. Results of specific individual tests should not be considered as a *prima facie* justification for determining validity. Validity assessment should include study-specific design and analysis procedures and outcomes, as well as consideration of knowledge from the body of preceding research.

Many design features to enhance validity and types of validity tests have been proposed in the literature, but no one type of validity test provides sufficient insight alone. Each study should hence include multiple validity tests targeted to specific issues or concerns that might arise within the application. The goal of testing should be to achieve a critical mass of evidence that supports, or not, the validity of estimated values.

When considering validity, it is important to recognize that bias does not, per se, preclude

validity. There is a bias/variance tradeoff with any type of estimate, and it may be impossible to purge an estimate of all bias. Under the strictest econometric interpretation, virtually all empirical estimates are subject to bias of some type. Validity ultimately relies on the procedures used to design and implement a study, and tests to identify and minimize potential sources of bias. The amount of bias that is tolerable, for the most part, is a matter of what is acceptable to support decision-making in specific contexts.

Finally, we emphasize that the validity or invalidity of a study cannot be ascertained by looking at individual responses, but instead by considering the sample of response data as a whole in the context of what is known in the literature. Likewise, the general validity or invalidity of SP methods (CV and CEs) cannot be ascertained from a single study, but from the weight of the evidence across the literature.

8. STUDY REPORTING

Study reporting is a fundamental element of the scientific method as it allows others to fully understand a study and provides the basis for replication. Effective communication is the foundation for future research to advance an empirical method and for the use and reuse of results from a given valuation study, e.g., for benefit transfers.

RECOMMENDATION 23: All studies, whether they are applied (to support decision-making) or methodological (to support the evolution of research), should fully document study design, implementation, analyses and results. Such transparency is crucial for the scientific credibility of studies and the appropriate interpretation and use of results. Documentation can also help support the use of the results to inform future decisions, even if this was not the initial intent of the study.

Study reporting and archival documentation is important for many reasons. For example, *ex post* content validity assessment of a study and efforts to replicate study results require documentation of procedures and investigator decisions (and related assumptions). The use of value estimates to support decision-making also requires documentation of study design, implementation, data coding, analyses and results. Finally, documentation is required so that future meta-analyses or systematic reviews can effectively control for differences across studies and applications (Loomis and Rosenberger 2006).

When implementing this recommendation, there may be tension between study documentation and journal page limits (and the patience of readers). While documentation of key study methods and results can be provided within the confines of journal articles, many journals now publish appendices online to provide study documentation, and technical documents can also be posted online (e.g., in data repositories or university archives). As we discuss in the next section, transparency is at the heart of credible scientific research. To meet this challenge, investigators should adhere to more detailed and systematic documentation of SP studies. Publication outlets should also provide a platform to support thorough study documentation.

9. CONCLUDING THOUGHTS AND COMMENTS

Writing a set of professional recommendations is easier said than done. Attention to balance is required so that the recommendations are clear, grounded in evidence from the literature and experience, and provide a basis for sound practice. For example, research innovations generally focus on advancing particular aspects of practice, and best-practice recommendations can establish an agreed-upon foundation for this work. Well-designed recommendations should also suggest opportunities for (and stimulate) innovations that enhance practice. This is particularly

important in cases where potential omissions, lack of clarity, or disagreement with proposed recommendations can provide incentives for research. These challenges are magnified in areas of methodology subject to controversy or disagreement. This paper seeks to raise the general quality of SP studies used to support decision-making, and to encourage a harmonization of expectations across those who use (or might consider using) SP methods to inform decisions. For areas where there has been no clear support for a particular practice, we indicate the need for future research. This is what we, as a team—and with broad solicitation of input from others in the profession and review of an extensive literature—have sought to accomplish.

The presented recommendations address major decisions in the design and implementation of an SP study. We recognize that there are details and considerations in implementing each practice, but we have chosen not to go into extensive detail regarding specific implementation decisions. With regard to specific design decisions, investigators must consider what the literature suggests, as well as issues that must be avoided or addressed. For example, we do not discuss in detail the selection and assignment of attribute levels in CEs nor bid levels in CV questions. Instead, we provide fundamental, though not exhaustive, references to assist readers in seeking more insight concerning specific implementation decisions that are not addressed by the overarching recommendations provided herein.

These recommendations emphasize the estimation of benefits and costs for changes to public goods and services; this is the primary application of SP methods within environmental and health economics. The incentive properties for valid estimates of value are becoming relatively well understood in such applications. SP questions are also applied to private goods and services (as they often are within transportation and marketing studies). However, the incentive properties of SP methods are less clear in those contexts. Further, in a private-good

application, some recommendations may be less reasonable, applicable or relevant. Thus, private-good applications can be informed by many of the recommendations provided, but should not be bound by these recommendations where they fail to apply in practice.

Although we have provided recommendations that are applicable to most contexts in which SP methods are applied, all studies must consider context-specific aspects of design and implementation. For example, a study seeking to estimate values for ecosystem service changes must consider issues such as the endpoints to be valued and how they are described (Boyd et al. 2016). Applications in both developed and developing countries must take into account unique cultural and institutional considerations (Bennett and Birol 2010). These examples once again highlight the need to (a) review the literature to learn from previous applications, (b) consult with local or context-specific experts, and (c) take care with design and pretesting.

SP studies draw, at a minimum, from four (and often five or more) disciplines. *Economics* guides the types of values to be estimated. *Survey research* provides the implementation procedures to collect data. *Statistics* (including econometrics) provides the tools to convert the data into estimated values. *Psychology* provides the basis for many insights into the theory and measurement of public perceptions and attitudes, causes and manifestations of different types of behavior (including behaviors considered to be anomalous within the context of economic theory), and the assessment of validity, among many other topics. Each of these fields has evolving sets of best practices necessary to produce credible scientific research, and these discipline-specific best practices are also worthy of consideration. Further, the *natural and health sciences* (and *engineering*) can inform the types of measures necessary to quantify baselines and the changes to be valued in an SP study (Boyd et al. 2016; Schultz et al. 2012).

We have discussed some of the ethical issues concerning data collection that involves

human subjects, but a discussion of broader ethical aspects of economic analysis is also required. As noted by the *Oxford Handbook of Professional Economic Ethics.* DeMartino and McCloskey (2016, p. 6) state "the case for economic ethics is simple and, we think, undeniable. Economists enjoy tremendous influence today over the life chances of others—innumerable others. That is the heart of the matter." This is particularly true for SP studies. While the results of these studies may not be the sole piece of information that guides a decision, this information is often a direct or indirect input into decisions that affect many people. Sources such as DeMartino and McCloskey (2016) provide important insights for applied economists and others designing and implementing any SP study, beyond the recommendations provided here.

In terms of data analyses, the American Statistical Association provides guidance to improve statistical analyses and reporting of the analyses (Wasserstein and Lazar 2016). Similar guidelines are provided by groups around the world, such as the Royal Statistical Society in the UK (http://www.rss.org.uk/, accessed October 4, 2016). Specific caution and insight is provided regarding the use and interpretation of p-values. One comment is particularly relevant to econometric analyses of SP data: "P-values can indicate how incompatible the data are with a specified statistical model" (Wasserstein and Lazar 2016, p. 131). Findings of statistical significance and insignificance are conditional on the assumed economic models, the manner in which the data have been collected, and the estimation methods used—any of which may or may not be appropriate. This guidance goes on to assert: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold" (p. 131), and "Proper inference requires full reporting and transparency" (p. 131). The predisposition of authors, reviewers and editors to favor results that pass accepted thresholds for statistical significance is a common source of publication bias in many research literatures,

including SP and RP studies (Rosenberger and Johnston 2009; Stanley and Doucouliagos 2012).

Hence, beyond the specific recommendations here, SP studies should adhere to general good practice in analysis, including transparency about the role of investigator-imposed assumptions. In a broad sense, the American Association for Public Opinion Research has a Transparency Initiative for survey research (http://www.aapor.org/transparency.aspx, accessed August 4, 2016). Transparency in reporting of SP studies—from study design and pretesting to implementation and data analysis—is crucial for value estimates to be credible.

Beyond the method-specific recommendations presented here, recognizing and following ethical and practice guidance provided by the supporting disciplines can enhance validity and reliability. This process can enhance the overall credibility of the outcome(s) of any single SP study as well as the method itself. Those who design and implement SP studies are encouraged to be familiar with the additional guidance available from related disciplines.

In closing, we offer novices *bon voyage*[78] in their early SP endeavors. For established practitioners, we hope the recommendations provide for *viages seguros*[79] in providing value estimates suitable to support decision-making. Researchers, you will find **具有挑**战性的努力[80] to enhance SP practice that support, refute or add to the recommendations provided here. We hope that the presented work encourages the use of SP studies to support decision-making and stimulates future research in this important area of inquiry.

[78] Have a nice trip (or voyage).
[79] Safe travels.
[80] Challenging endeavors.

REFERENCES

Aadland, David, and Arthur J. Caplan. 2006. Cheap talk reconsidered: New evidence from CVM. *Journal of Economic Behavior & Organization* 60:562-78.

Adamowicz, Wiktor, Jordan Louviere, and Michael Williams. 1994. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management* 26:271-92.

Adamowicz, Wiktor, Joffre Swait, Peter Boxall, Jordan Louviere, and Michael Williams. 1997. Perceptions versus objective measures of environmental quality in combined revealed and stated preference models of environmental valuation. *Journal of Environmental Economics and Management* 32:65-84.

Adamowicz, Wiktor, Peter Boxall, Michael Williams, and Jordan Louviere. 1998. Stated preference approaches for measuring passive use values: Choice experiments and contingent valuation. *American Journal of Agricultural Economics* 80:64-75.

Adamowicz, Wiktor, Mark Dickie, Shelby Gerking, Marcella Veronesi, and David Zinner. 2014. Household decision making and valuation of environmental health risks to parents and their children. *Journal of the Association of Environmental and Resource Economists* 1:481-519.

Ajzen, Icek, Thomas C. Brown, and Lori H. Rosenthal. 1996. Information bias in contingent valuation: Effects of personal relevance, quality of information, and motivational orientation. *Journal of Environmental Economics and Management* 30:43-57.

Akter, Sonia, Jeff Bennett, and Michael B. Ward. 2012. Climate change, scepticism and public support for mitigation: Evidence from an Australian choice experiment. *Global Environmental Change* 22:736-45.

Alberini, Anna. 1995. Optimal designs for discrete choice contingent valuation surveys: Single-bound, double-bound, and bivariate models. *Journal of Environmental Economics and Management* 28:287-306.

Alemu, Mohammed Hussen, Morten Raun Mørkbak, Søren Bøye Olsen, and Carsten Lynge Jensen. 2013. Attending to the reasons for attribute non-attendance in choice experiments. *Environmental and Resource Economics* 54:333-59.

Alevy, Jonathan E., John A. List, and Wiktor L. Adamowicz. 2011. How can behavioral economics inform nonmarket valuation? An example from the preference reversal literature. *Land Economics* 87:365-81.

Alston, Richard M., and Clifford Nowell. 1996. Implementing the voluntary contribution game: A field experiment. *Journal of Economic Behavior & Organization* 31:357-68.

Andreoni, James. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97:1447-58.

Angrist, Joshua D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics* 19:2-28.

Araña, Jorge E. and Carmelo J. León. 2008. Do emotions matter? Coherent preferences under anchoring and emotional effects. *Ecological Economics* 66:700-11.

Araña, Jorge E., Carmelo J. León, and Michael W. Hanemann. 2008. Emotions and decision rules in discrete choice experiments for valuing health care programmes for the elderly. *Journal of Health Economics* 27:753-69.

Arentze, Theo, Aloys Borgers, Harry Timmermans, and Romano DelMistro. 2003. Transport stated choice responses: Effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review* 39:229-44.

Arrow, Kenneth, Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, and Howard Schuman. 1993. Report of the NOAA panel on contingent valuation. *Federal Register* 58:4601-14.

Atkinson, Giles, Sian-Morse Jones, Susana Mourato, and Allan Provins. 2012. When to take "no" for an answer? Using entreaties to reduce protests in contingent valuation studies. *Environmental and Resource Economics* 51:497-523.

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Cooper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau. 2013. Summary report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology* 1:90-143.

Baker, Rick, and Brad Ruting. 2014. *Environmental policy analysis: A guide to non-market valuation.* Melbourne, Australia: Productivity Commission.

Balcombe, Kelvin, and Iain Fraser. 2011. A general treatment of 'don't know' responses from choice experiments. *European Review of Agricultural Economics* 38:171-91.

Balcombe, Kelvin, Ali Chalak, and Iain Fraser. 2009. Model selection for the mixed logit with Bayesian estimation. *Journal of Environmental Economics and Management* 57:226-37.

Balcombe, Kelvin, Michael Burton, and Dan Rigby. 2011. Skew and attribute non-attendance within the Bayesian mixed logit model. *Journal of Environmental Economics and Management* 62: 446-61.

Balcombe, Kelvin, Iain Fraser, and Eugene McSorley. 2015. Visual attention and attribute attendance in multi-attribute choice experiments. *Journal of Applied Econometrics* 30:447-67.

Banzhaf, H. Spencer, Dallas Burtraw, David Evans, and Alan Krupnick. 2006. Valuation of natural resource improvements in the Adirondacks. *Land Economics* 82:445-64.

Barrios, Javier A. 2004. Generalized sample selection bias correction under RUM. *Economics Letters* 85:129-32.

Bateman, Ian J., and Ian H. Langford. 1997. Budget-constraint, temporal, and question-ordering effects in contingent valuation studies. *Environment and Planning A* 29:1215-28.

Bateman, Ian J., and Kenneth G. Willis, eds. 1999. *Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries.* New York, NY and Oxford, UK: Oxford University Press.

Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer, and Robert Sugden. 1997a. Does part-whole bias exist? An experimental investigation. *The Economic Journal* 107:322-32.

Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer, and Robert Sugden. 1997b. A test of the theory of reference-dependent preferences. *The Quarterly Journal of Economics* 112:479-505.

Bateman, Ian J., Ian H. Langford, Andrew P. Jones, and Geoffrey N. Kerr. 2001. Bound and path effects in double and triple bounded dichotomous choice contingent valuation. *Resource and Energy Economics* 23:191-213.

Bateman, Ian J., Richard T. Carson, Brett H. Day, W. Michael Hanemann, Nick Hanley, Tanis Hett, Michael Jones-Lee, Graham Loomes, Susana Mourato, Ece Özdemiroglu, and David W. Pearce. 2002. *Economic valuation with stated preference techniques: A manual*. Cheltenham, UK: Edward Elgar.

Bateman, Ian J., Matthew Cole, Philip Cooper, Stavros Georgiou, David Hadley, and Gregory L. Poe. 2004. On visible choice sets and scope sensitivity. *Journal of Environmental Economics and Management* 47:71-93.

Bateman, Ian J., Philip Cooper, Stavros Georgiou, Ståle Navrud, Gregory L. Poe, Richard Ready, Pere Riera, Mandy Ryan, and Christian A. Vossler. 2005. Economic valuation of policies for managing acidity in remote mountain lakes: Examining validity through scope sensitivity testing. *Aquatic Sciences* 67:274-91.

Bateman, Ian J., Brett H. Day, Stavros Georgiou, and Iain Lake. 2006a. The aggregation of environmental benefit values: Welfare measures, distance decay and total WTP. *Ecological Economics* 60:450-60.

Bateman, Ian J., Alistair Munro, Bruce Rhodes, Chris V. Starmer, and Robert Sugden. 2006b. Anchoring and yea-saying with private goods: An experiment. In *Using experimental methods in environmental and resource economics*, ed. John A. List, 1-19. Cheltenham, UK: Edward Elgar.

Bateman, Ian J., Diane Burgess, W. George Hutchinson, and David I. Matthews. 2008a. Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness. *Journal of Environmental Economics and Management* 55:127-41.

Bateman, Ian J., Alistair Munro, and Gregory L. Poe. 2008b. Decoy effects in choice experiments contingent valuation: Asymmetric dominance. *Land Economics* 84:115-27.

Bateman, Ian J., Brett H. Day, Diane P. Dupont, and Stavros Georgiou. 2009. Procedural invariance testing of the one-and-one-half-bound dichotomous choice elicitation method. *The Review of Economics and Statistics* 91:806-20.

Bech, Mickael, Trine Kjaer, and Jørgen Lauridsen. 2011. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. *Health Economics* 20:273-86.

Bell, Jason, Joel Huber, and W. Kip Viscusi. 2011. Survey mode effects on valuation of environmental goods. *Int J Environ Res Public Health* 8:1222-43

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 2005. *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, NJ: John Wiley & Sons.

Ben-Akiva, Moshe, Daniel McFadden, Tommy Gärling, Dinesh Gopinath, Joan Walker, Denis Bolduc, Axel Börsch-Supan, Philippe Delquié, Oleg Larichev, Taka Morikawa, Amalia Polydoropoulou, and Vithala Rao. 1999. Extended Framework for Modeling Choice Behavior. *Marketing Letters* 10:187-203.

Bennett, Jeff. 1996. The contingent valuation: A post Kakadu assessment. *Agenda: A Journal of Policy Analysis and Reform* 3:185-94.

Bennett, Jeff, and Ekin Birol, eds. 2010. *Choice experiments in developing countries*, Cheltenham, UK: Edward Elgar.

Bennett, Jeff, and Russell Blamey, eds. 2001. *The choice modelling approach to environmental valuation*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Bergstrom, John C., John R. Stoll, and Alan Randall. 1989. Information effects in contingent markets. *American Journal of Agricultural Economics* 71: 685-91.

Bishop, Richard C., and Kevin J. Boyle. 2017. Reliability and validity in nonmarket valuation. In *A primer on nonmarket valuation*, eds. Patricia A. Champ, Kevin J. Boyle, and Thomas C. Brown. Netherlands: Springer Science and Business Media.

Bishop, Richard C., and Thomas A. Heberlein. 1979. Measuring values of extramarket goods: Are indirect measures biased? *American Journal of Agricultural Economics* 61:926-30.

Blamey, Russell K., Jeff W. Bennett, and Mark D. Morrison. 1999. Yea-saying in contingent valuation surveys. *Land Economics* 75:126-41.

Bliem, Marcus, Michael Getzner, and Petra Rodiga-Laßnig. 2012. Temporal stability of individual preferences for river restoration in Austria using a choice experiment. *Journal of Environmental Management* 103:65-73.

Bliemer, Michiel C.J., and John M. Rose. 2010. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological* 44:720-34.

Blomquist, Glenn C., Karen Blumenschein and Magnus Johannesson. 2009. Eliciting willingness to pay without bias using follow-up certainty statements: Comparisons between probably/definitely and a 10-point certainty scale. *Environmental and Resource Economics* 43:473-502.

Blomquist, Glenn C., and John C. Whitehead. 1998. Resource quality information and validity of willingness to pay in contingent valuation. *Resource and Energy Economics* 20:179-96.

Blumberg, Stephen J., Julian V. Luke, Nadarajasundaram Ganesh, Michal E. Davern, and Michael H. Boudreaux. 2012. Wireless substitution: State-level estimates from the National Health Interview Survey, 2010-2011. National Health Statistics Reports Number 61. Hyattsville, MD: Centers for Disease Control and Prevention.

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman. 2008. Eliciting willingness to pay without bias: Evidence from a field experiment. *The Economic Journal* 118:114-37.

Boeri, Marco, Riccardo Scarpa, and Caspar G. Chorus. 2014. Stated choices and benefit estimates in the context of traffic calming schemes: Utility maximization, regret minimization, or both? *Transportation Research Part A: Policy and Practice* 61:121-35.

Bonnieux, Francois, and Pierre Rainelli. 1999. Contingent valuation methodology and the EU institutional framework. In *Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries*, ed. Ian J. Bateman, and Kenneth George Willis. Oxford: Oxford University Press.

Bosworth, Ryan, Trudy Ann Cameron, and J.R. DeShazo. 2010. Is an ounce of prevention worth a pound of cure? Comparing demand for public prevention and treatment policies. *Medical Decision Making* 30:E40-E56.

Botzen, W.J. Wouter, and Jeroen C.J.M. van den Bergh. 2012. Monetary valuation of insurance against flood risk under climate change. *International Economic Review* 53:1005-25.

Boxall, Peter C., Wiktor L. Adamowicz, and Amanda Moon. 2009. Complexity in choice experiments: Choice of the status quo alternative and implications for welfare measurement. *Australian Journal of Agricultural and Resource Economics* 53:503-19.

Boxall, Peter C., Wiktor L. Adamowicz, Maria Olar, Gale E. West, and Guy Cantin. 2012. Analysis of the economic benefits associated with the recovery of threatened marine mammal species in the Canadian St. Lawrence Estuary. *Marine Policy* 36:189-97.

Boxall, Peter C., Wiktor L. Adamowicz, Joffre Swait, Michael Williams, and Jordan Louviere. 1996. A comparison of stated preference methods for environmental valuation. *Ecological Economics* 18:243-53.

Boyd, James, and Alan Krupnick. 2009. The definition and choice of environmental commodities for nonmarket valuation. RFF Discussion Paper 09-35, Resources for the Future, Washington, DC.

———. 2013. Using ecological production theory to define and select environmental commodities for nonmarket valuation. *Agricultural and Resource Economics Review* 42:1-32.

Boyd, James, Paul Ringold, Alan Krupnick, Robert J. Johnston, Matthew A. Weber, and Kim Hall. 2016. Ecosystem services indicators: Improving the linkage between biophysical and economic analyses. *International Review of Environmental and Resource Economics* 8:359-443.

Boyle, Kevin J. 1989. Commodity specification and the framing of contingent-valuation questions. *Land Economics* 65:57-63.

Boyle, Kevin J., and Richard C. Bishop. 1988. Welfare measurements using contingent valuation: A comparison of techniques. *American Journal of Agricultural Economics* 70:20-8.

Boyle, Kevin J., Richard C. Bishop, and Michael P. Welsh. 1985. Starting point bias in contingent valuation bidding games. *Land Economics* 61:188-94.

———. 1993. The role of question order and respondent experience in contingent-valuation studies. *Journal of Environmental Economics and Management* 25:S80-S99.

Boyle, Kevin J., F. Reed Johnson, and Daniel W. McCollum. 1997. Anchoring and adjustment in single-bounded, contingent-valuation questions. *American Journal of Agricultural Economics* 9:1495-500.

Boyle, Kevin J., F. Reed Johnson, Daniel W. McCollum, William H. Desvousges, Richard W. Dunford, and Sara P. Hudson. 1996. Valuing public goods: Discrete versus continuous contingent-valuation responses. *Land Economics* 72:381-96.

Boyle, Kevin J., Thomas P. Holmes, Mario F. Tiesl, and Brian Roe. 2001. A comparison of conjoint analysis response formats. *American Journal of Agricultural Economics* 83:441-54.

Boyle, Kevin J., Nicolai V. Kuminoff, Christopher F. Parmeter, and Jaren C. Pope. 2010. The benefit-transfer challenges. *Annual Review of Resource Economics* 2:161-82.

Boyle, Kevin J., Mark Morrison, Darla Hatton MacDonald, Roderick Duncan, and John Rose. 2016. Investigating internet and mail implementation of stated-preference surveys while controlling for differences in sample frames. *Environmental and Resource Economics* 64:401-19.

Boyle, Kevin J., and Semra Özdemir. 2009. Convergent validity of attribute-based, choice questions in stated-preference studies. *Environmental and Resource Economics* 42:247-64.

Braga, Jacinto, and Chris Starmer. 2005. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics* 32:55-89.

Brick, J. Michael, and Douglas Williams. 2013. Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science* 645:36-59.

Bridges, John F.P., A. Brett Hauber, Deborah Marshall, Andrew Lloyd, Lisa A. Prosser, Dean A. Regier, F. Reed Johnson, and Josephine Mauskopf. 2011. Conjoint analysis application in health—a checklist: A report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in Health* 14:403-13.

Brouwer, Roy, Ian H. Langford, Ian J. Bateman, and R. Kerry Turner. 1999. A meta-analysis of wetland contingent valuation studies. *Regional Environmental Change* 1:47-57.

Brouwer, Roy, Thijs Dekker, John Rolfe, and Jill Windle. 2010a. Choice certainty and consistency in repeated choice experiments. *Environmental and Resource Economics* 46:93-109.

Brouwer, Roy, Julia Martin-Ortega, and Julio Berbel. 2010b. Spatial preference heterogeneity: A choice experiment. *Land Economics* 86:552-68.

Brouwer, Roy, and Julia Martín-Ortega. 2012. Modeling self-censoring of polluter pays protest votes in stated preference research to support resource damage estimations in environmental liability. *Resource and Energy Economics* 34:151-66.

Brown, Thomas C., and Robin Gregory. 1999. Why the WTA-WTP disparity matters. *Ecological Economics* 28:323-35.

Bujosa, Angel, Antoni Riera, and Robert L. Hicks. 2010. Combining discrete and continuous representations of preference heterogeneity: A latent class approach. *Environmental and Resource Economics* 47:477-93.

Bulte, Erwin, Shelby Gerking, John A. List, and Aart de Zeeuw. 2005. The effect of varying the cause of environmental problems on stated WTP values: Evidence from a field study. *Journal of Environmental Economics and Management* 49:330-42.

Burghart, Daniel R., Trudy Ann Cameron, and Geoffrey A. Gerdes. 2007. Valuing publicly sponsored research projects: Risks, scenario adjustments, and inattention. *Journal of Risk and Uncertainty* 35:77-105.

Burr, Jennifer M., Paola Botello-Pinzon, Yemisi Takwoingi, Rodolfo Hernández, Maria Vazquez-Montes, Andrew Elders, Ryo Asaoka, Kathryn Banister, Josine van der Schoot, Cynthia Fraser, Anthony King, Hans Lemij, Roshini Sanders, Stephen Vernon, Anja Tuulonen, Aachal Kotecha, Paul Glasziou, David Garway-Heath, David Crabb, Luke Vale, Augusto Azuara-Blanco, Rafael Perera, Mandy Ryan, Jon Deeks, and Jonathan Cook. 2012. Surveillance for ocular hypertension: An evidence synthesis and economic evaluation. *Health Technology Assessment* 16:1-271.

Burton, Michael, and Dan Rigby. 2012. The self selection of complexity in choice experiments. *American Journal of Agricultural Economics* 94:786-800.

Bush, Glenn, Nick Hanley, Mirko Moro, and Daniel Rondeau. 2013. Measuring the local costs of conservation: A provision point mechanism for eliciting willingness to accept compensation. *Land Economics* 89:490-513.

Cai, Beilei, Trudy Ann Cameron, and Geoffrey R. Gerdes. 2010. Distributional preferences and the incidence of costs and benefits in climate change policy. *Environmental and Resource Economics* 46:429-58.

Cameron, Trudy Ann. 2005a. Individual option prices for climate change mitigation. *Journal of Public Economics* 89:283-301.

Cameron, Trudy Ann. 2005b. Updating subjective risks in the presence of conflicting information: An application to climate change. *Journal of Risk and Uncertainty* 30:63-97.

Cameron, Trudy Ann, and J.R. DeShazo. 2005. Comprehensive selectivity assessment for a major consumer panel: Attitudes toward government regulation of environment, health, and safety risks. Unpublished paper, Department of Economics, University of Oregon.

Cameron, Trudy Ann, and J.R. DeShazo. 2010. Differential attention to attributes in utility-theoretic choice models. *Journal of Choice Modelling* 3:73-115.

Cameron, Trudy Ann, and J.R. DeShazo. 2013. Demand for health risk reductions. *Journal of Environmental Economics and Management* 65:87-109.

Cameron, Trudy Ann, and Jeffrey Englin. 1997. Respondent experience and contingent valuation of environmental goods. *Journal of Environmental Economics and Management* 33:296-313.

Cameron, Trudy Ann, and Michelle D. James. 1987. Efficient estimation methods for "closed-ended" contingent valuation surveys. *The Review of Economics and Statistics* 69:269-76.

Cameron, Trudy Ann, and John Quiggin. 1994. Estimation using contingent valuation data from a "dichotomous choice with follow-up" questionnaire. *Journal of Environmental Economics and Management* 27:218-34.

Cameron, Trudy Ann, J.R. DeShazo, and Erica H. Johnson. 2011. Scenario adjustment in stated preference research. *Journal of Choice Modelling* 4:9-43.

Cameron, Trudy Ann, Gregory L. Poe, Robert G. Ethier, and William D. Schulze. 2002. Alternative non-market value-elicitation methods: Are the underlying preferences the same? *Journal of Environmental Economics and Management* 44:391-425.

Campbell, Danny, W. George Hutchinson, and Riccardo Scarpa. 2009. Using choice experiments to explore the spatial distribution of willingness to pay for rural landscape improvements. *Environment and Planning A* 41:97-111.

———. 2008a. Incorporating discontinuous preferences into the analysis of discrete choice experiments. *Environmental and Resource Economics* 41:401-17.

Campbell, Danny, David A. Hensher, and Riccardo Scarpa. 2011. Non-attendance to attributes in environmental choice analysis: A latent class specification. *Journal of Environmental Planning and Management* 54:1061-76.

Campbell, Danny, Riccardo Scarpa, and W. George Hutchinson. 2008b. Assessing the spatial dependence of welfare estimates obtained from discrete choice experiments. *Letters in Spatial and Resource Sciences* 1: 117-26.

Campos, Pablo, Alejandro Caparros, and Jose L. Oviedo. 2007. Comparing payment-vehicle effects in contingent valuation studies for recreational use in two protected Spanish forests. *Journal of Leisure Research* 39:60-85.

Carlsson, Fredrik, and Peter Martinsson. 2003. Design techniques for stated preference methods in health economics. *Health Economics* 12:281-94.

Carlsson, Fredrik, Morten Raun Mørkbak, and Søren Bøye Olsen. 2012. The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling* 5:19-37.

Carlsson, Fredrik, Mitesh Kataria, Alan Krupnick, Elina Lampi, Åsa Löfgren, Ping Qin, and Thomas Sterner. 2013. The truth, the whole truth, and nothing but the truth—A multiple country test of an oath script. *Journal of Economic Behavior & Organization* 89:105-21.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and validity assessment.* Volume 17 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Woburn, MA: Sage.

Carrasco, Raquel. 2012. Binary choice with binary endogenous regressors in panel data: Estimating the effect of fertility on female labor participation. *Journal of Business and Economic Statistics* 19:385-94.

Carson, Katherine S., Susan M. Chilton, and W. George Hutchinson. 2009. Necessary conditions for demand revelation in double referenda. *Journal of Environmental Economics and Management* 57:219-25.

Carson, Richard T. 1997. Contingent valuation surveys and tests of insensitivity to scope. In *Determining the value of non-marketed goods*, ed. Raymond J. Kopp, Werner W. Pommerehne, and Norbert Schwarz, 127-63. Boston: Kluwer Academic Publishers.

Carson, Richard T. 2000. Contingent valuation: A user's guide. *Environmental Science & Technology* 34:1413-18.

Carson, Richard T. 2011. Contingent *valuation*: A *comprehensive bibliography* and *history*. Cheltenham, UK: Edward Elgar.

Carson, Richard T. 2012. Contingent valuation: A practical alternative when prices aren't available. *The Journal of Economic Perspectives* 26:27-42.

Carson, Richard T., and Theodore Groves. 2007. Incentive and informational properties of preference questions. *Environmental and Resource Economics* 37:181-210.

Carson, Richard T., and W. Michael Hanemann. 2005. Contingent valuation. *Handbook of Environmental Economics* 2:821-936.

Carson, Richard T. and Robert Cameron Mitchell. 1993. The issue of scope in contingent valuation studies. *American Journal of Agricultural Economics* 75:1263-67.

Carson, Richard T., and Robert Cameron Mitchell. 1995. Sequencing and nesting in contingent valuation surveys. *Journal of Environmental Economics and Management* 28:155-73.

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* 72:80-99.

Carson, Richard T., Nicholas E. Flores, and W. Michael Hanemann. 1998a. Sequencing and valuing public goods. *Journal of Environmental Economics and Management* 36:314-23.

Carson, Richard T., W. Michael Hanemann, Raymond J. Kopp, Jon A. Krosnick, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith. 1994. Prospective interim lost use value due to PCB and DDT contamination in the Southern California Bight: Volume II (Appendices). La Jolla, CA: US Department of Commerce (NOAA).

Carson, Richard T., W. Michael Hanemann, Raymond J. Kopp, Jon A. Krosnick, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith, with Michael Conaway, and Kerry Martin. 1998b. Referendum design and contingent valuation: The NOAA panel's no-vote recommendation. *The Review of Economics and Statistics* 80:335-38.

Carson, Richard T., Nicholas E. Flores, and Norman F. Meade. 2001. Contingent valuation: Controversies and evidence. *Environmental and Resource Economics* 19:173-210.

Carson, Richard T., Theodore Groves, and John A. List. 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice. *Journal of the Association of Environmental and Resource Economists* 1:171-207.

Carson, Richard T., Robert C. Mitchell, Michael Hanemann, Raymond J. Kopp, Stanley Presser, and Paul A. Ruud. 2003. Contingent valuation and lost passive use: Damages from the Exxon Valdez oil spill. *Environmental and Resource Economics* 25:257-89.

Carson, Richard T., and Jordan J. Louviere. 2011. A common nomenclature for stated preference elicitation approaches. *Environmental and Resource Economics* 49:539-59.

Caussade, Sebastian, Juan de Dios Ortúzar, Luis I. Rizzi, and David A. Hensher. 2005. Assessing the influence of design dimensions on stated choice experiments. *Transportation Research Part B: Methodological* 39:621-40.

Champ, Patricia A., and Richard C. Bishop. 2001. Donation payment mechanisms and contingent valuation: An empirical study of hypothetical bias. *Environmental and Resource Economics* 19:383-402.

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* 33:151-62.

Champ, Patricia A., Kevin C. Boyle, and Thomas C. Brown, eds. 2003. *A primer on nonmarket valuation.* New York: Springer.

Champ, Patricia A., Kevin C. Boyle, and Thomas C. Brown, eds. 2017. *A primer on nonmarket valuation*. Netherlands: Springer Science and Business Media.

Champ, Patricia A., Rebecca Moore, and Richard C. Bishop. 2009. A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review* 38:166-80.

Chilton, Susan M., and W. George Hutchinson. 1999. Do focus groups contribute anything to the contingent valuation process? *Journal of Economic Psychology* 20:465-83.

Chrzan, Keith. 1994. Three kinds of order effects in choice-based conjoint analysis. *Marketing Letters* 5:165-72.

Ciriacy-Wantrup, Siegfried V. 1947. Capital returns from soil-conservation practices. *Journal of Farm Economics* 29:1181-96.

Clarke, Philip M. 2002. Testing the convergent validity of the contingent valuation and travel cost methods in valuing the benefits of health care. *Health Economics* 11:117-27.

Coast, Joanna. 1999. The appropriate uses of qualitative methods in health economics. *Health Economics* 8:345-53.

Coast, Joanna, Ruth McDonald, and Rachel Baker. 2004. Issues arising from the use of qualitative methods in health economics. *Journal of Health Services Research & Policy* 9:171-6.

Coast, Joanna, and Susan A. Horrocks. 2007. Developing attributes and levels for discrete choice experiments using qualitative methods. *Journal of Health Services Research & Policy* 12:25-30.

Coast, Joanna, Hareth Al-Janabi, Eileen J. Sutton, Susan A. Horrocks, A. Jane Vosper, Dawn R. Swancutt, and Terry N. Flynn. 2012. Using qualitative methods for attribute development for discrete choice experiments: Issues and recommendations. *Health Economics* 21:730-41.

Collins, Jill P., and Christian A. Vossler. 2009. Incentive compatibility tests of choice experiment value elicitation questions. *Journal of Environmental Economics and Management* 58:226-35.

Colson, Gregory, Jay R. Corrigan, Carola Grebitus, Maria L. Loureiro, and Matthew C. Rousu. 2016. Which deceptive practices, if any, should be allowed in experimental economics research? Results from surveys of applied experimental economists and students. *American Journal of Agricultural Economics* 98:610-21.

Cooper, Joseph C. 1993. Optimal bid selection for dichotomous choice contingent valuation surveys. *Journal of Environmental Economics and Management* 24:25-40.

Cooper, Joseph C., and W. Michael Hanemann. 1994. Referendum contingent valuation: How many bounds are enough? *American Journal of Agricultural Economics* 76:1246.

Cooper, Joseph C., W. Michael Hanemann, and Giovanni Signorello. 2002. One-and-one-half-bound dichotomous choice contingent valuation. *The Review of Economics and Statistics* 84:742-50.

Cooper, Joseph C., and John Loomis. 1992. Sensitivity of willingness-to-pay estimates to bid design in dichotomous choice contingent valuation models. *Land Economics* 68:211-224.

Corso, Phaedra S., James K. Hammitt, and John D. Graham. 2001. Valuing mortality-risk reduction: Using visual aids to improve the validity of contingent valuation. *Journal of Risk and Uncertainty* 23:165-84.

Covey, Judith, Graham Loomes, and Ian J. Bateman. 2007. Valuing risk reductions: Testing for range biases in payment card and random card sorting methods. *Journal of Environmental Planning and Management* 50:467-82.

Craig, Benjamin M., Shannon K. Runge, Kim Rand-Hendriksen, Juan Manuel Ramos-Goñi, and Mark Oppe. 2015. Learning and satisficing: An analysis of sequence effects in health valuation. *Value in Health* 18:217-23.

Croson, Rachel. 2003. Why and how to experiment: Methodologies from experimental economics. *University of Illinois Law Review* (January 2002):921-45.

Cummings, Ronald G., and Laura O. Taylor. 1999. Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *The American Economic Review* 89:649-65.

Cummings, Ronald G., David S. Brookshire, William D. Schulze, Richard C. Bishop, and Kenneth Joseph Arrow. 1986. *Valuing environmental goods: An assessment of the contingent valuation method.* Totowa, NJ: Rowman & Allanheld.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. The effects of response rate changes on the Index of Consumer Sentiment. *Public Opinion Quarterly* 64:413-28.

Czajkowski, Mikołaj, Nick Hanley, and Karine Nyborg. 2015. Social norms, morals and self-interest as determinants of pro-environment behaviours: The case of household recycling. *Environmental and Resource Economics*. Online First, DOI 10.1007/s10640-015-9964-3.

Daly, Andrew, Stephane Hess, and Kenneth Train. 2012. Assuring finite moments for willingness to pay in random coefficient models. *Transportation* 39:19-31.

Davis, Robert Kenneth. 1963. The value of outdoor recreation: An economic study of the Maine woods. PhD diss., Harvard University, Department of Economics.

Day, Brett, and Jose-Luis Prades. 2010. Ordering anomalies in choice experiments. *Journal of Environmental Economics and Management* 59:271-85.

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* 63:73-91.

deBekker-Grob, Esther W., Mandy Ryan, and Karen Gerard. 2012. Discrete choice experiments in health economics: A review of the literature. *Health Economics* 21:145-72.

Dekker, Thijs, Stephane Hess, Roy Brouwer, and Marjan Hofkes. 2016. Decision uncertainty in multi-attribute stated preference studies. *Resource and Energy Economics* 43:57-73.

De Leeuw, Edith, Mario Callegaro, Joop Hox, Elly Korendijk, and Gerty Lensvelt-Mulders. 2007. The influence of advance letters on response in telephone surveys: A meta-analysis. *Public Opinion Quarterly* 71:413-43.

Dellaert, Benedict G.C., Bas Donkers, and Arthur Van Soest. 2012. Complexity effects in choice experiment-based models. *Journal of Marketing Research* 49:424-34.

DeMartino, George F., and Deirdre N. McCloskey, eds. 2016. *The Oxford handbook of professional economic ethics.* New York: Oxford University Press.

DeShazo, J.R., and German Fermo. 2002. Designing choice sets for stated preference methods: The effects of complexity on choice consistency. *Journal of Environmental Economics and Management* 44:123-43.

Desvousges, William H., and James H. Frey. 1989. Integrating focus groups and surveys: Examples from environmental risk studies. *Journal of Official Statistics* 5:349-63.

Desvousges, William H., and V. Kerry Smith. 1988. Focus groups and risk communication: The "science" of listening to data. *Risk Analysis* 8:479-84.

Desvousges, William H., V. Kerry Smith, Diane H. Brown, and D. Kirk Pate. 1984. The role of focus groups in designing a contingent valuation survey to measure the benefits of hazardous waste management regulations. Final report to the US Environmental Protection Agency under Contract No. 68-01-6595. Research Triangle Park, NC: Research Triangle Institute.

Desvouges, William H., Kristy Mathews, and Kenneth Train. 2012. Adequate responsiveness to scope in contingent valuation. *Ecological Economics* 84:121-28.

———. 2015. An adding-up test on contingent valuations of river and lake quality. *Land Economics* 91:556-71.

Diamond, Peter. 1996. Testing the internal consistency of contingent valuation surveys. *Journal of Environmental Economics and Management* 30:337-47.

Diamond, Peter A., and Jerry A. Hausman. 1994. Contingent valuation: Is some number better than no number? *The Journal of Economic Perspectives* 8:45-64.

DiClemente, Diane F., and Donald A. Hantula. 2003. Applied behavioral economics and consumer choice. *Journal of Economic Psychology* 24:589-602.

Dillman, Don A. 1978. *Mail and telephone surveys: The total design method*. Hoboken, NJ: Wiley.

Dillman, Don A. 1991. The design and administration of mail surveys. *Annual Review of Sociology* 17:225-49.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, phone, mail and mixed-mode surveys: The tailored design method.* Hoboken, NJ: Wiley.

Edwards, Steven F., and Glen D. Anderson. 1987. Overlooked biases in contingent valuation surveys: Some considerations. *Land Economics* 63:168-78.

Egan, Kevin J., Jay R. Corrigan, and Daryl F. Dwyer. 2015. Three reasons to use annual payments in contingent valuation surveys: Convergent validity, discount rates, and mental accounting. *Journal of Environmental Economics and Management* 72:123-36.

Ericson, Keith M. Marzilli, and Andreas Fuster. 2014. The endowment effect. *Annual Review of Economics* 6:555-79.

Farrell, Joseph, and Robert Gibbons. 1989. Cheap talk can matter in bargaining. *Journal of Economic Theory* 48:221-37.

Feick, Lawrence F. 1989. Latent class analyses of survey questions that include don't know responses. *Public Opinion Quarterly* 53:525-47.

Ferrini, Silvia, and Riccardo Scarpa. 2007. Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management* 53:342-63.

Fiebig, Denzel G., Michael P. Keane, Jordan Louviere, and Nada Wasi. 2010. The generalized multinomial logit model: Accounting for scale and heterogeneity. *Marketing Science* 29:393-421.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, and Katherine Baicker. 2012. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics* 127: 1057-1106.

Fischhoff, Baruch, Noel T. Brewer, and Julia S. Downs, eds. 2011. *Communicating risks and benefits: An evidence-based user's guide*. Silver Spring, MD: Food and Drug Administration, US Department of Health and Human Services.

Flynn, Terry N. 2010. Valuing citizen and patient preferences in health: Recent developments in three types of best-worst scaling. *Expert Review of Pharmacoeconomics & Outcomes Research* 10:259-67.

Foster, Vivien, and Susana Mourato. 2003. Elicitation format and sensitivity to scope. *Environmental and Resource Economics* 24:141-60.

Freeman, A. Myrick III, Joseph A. Herriges, and Catherine L. Kling. 2014. *The measurement of environmental and resource values: Theory and methods.* New York: Routledge.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group. 1999. *Simple heuristics that make us smart.* New York: Oxford University Press.

Glenk, Klaus, and Sergio Colombo. 2011. How sure can you be? A framework for considering delivery uncertainty in benefit assessments based on stated preference methods. *Journal of Agricultural Economics* 62:25-46.

———. 2013. Modeling outcome-related risk in choice experiments. *Australian Journal of Agricultural and Resource Economics* 57:559-78.

Goldberg, Isabel, and Jutta Roosen. 2007. Scope insensitivity in health risk reduction studies: A comparison of choice experiments and the contingent valuation method for valuing safer food. *Journal of Risk and Uncertainty* 34:123-44.

Graham, Daniel A. 1981. Cost-benefit analysis under uncertainty. *The American Economic Review* 71:715-25.

Green, Donald, Karen E. Jacowitz, Daniel Kahneman, and Daniel McFadden. 1998. Referendum contingent valuation, anchoring, and willingness to pay for public goods. *Resource and Energy Economics* 20:85-116.

Green, Paul E., and Vithala R. Rao. 1971. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research* 8:355-63.

Green, Paul E., and V. Srinivasan. 1978. Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research* 5:103-23.

Greene, William H., David A. Hensher, and John M. Rose. 2005. Using classical simulation-based estimators to estimate individual WTP values. In *Applications of simulation methods in environmental and resource economics,* eds. Riccardo Scarpa and Anna Alberini. New York: Springer.

Groothuis, Peter A., and John C. Whitehead. 2002. Does don't know mean no? Analysis of "don't know" responses in dichotomous choice contingent valuation questions. *Applied Economics* 34:1935-40.

Groves, Robert M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* 70:646-75.

Groves, Robert M., and Steven G. Heeringa. 2006. Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169:439-57.

Groves, Robert M., and Emilia Peytcheva. 2008. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly* 72:167-89.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey methodology*. Hoboken, NJ: Wiley.

Groves, Theodore, Roy Radner, and Stanley Reiter, eds. 1987. *Information, incentives, and economic mechanisms: Essays in honor of Leonid Hurwicz.* Minneapolis: University of Minnesota Press.

Haab, Timothy C., and Kenneth McConnell. 2002. *Valuing environmental and natural resources: The econometrics of non-market valuation.* Cheltenham, UK: Edward Elgar.

Haab, Timothy C., Ju-Chin Huang, and John C. Whitehead. 1999. Are hypothetical referenda incentive compatible? A comment. *Journal of Political Economy* 107:186-96.

Haab, Timothy C., Matthew G. Interis, Daniel R. Petrolia, and John C. Whitehead. 2013. From hopeless to curious? Thoughts on Thomas Hausman's "dubious to hopeless" critique of contingent valuation. *Applied Economic Perspectives and Policy* 35:593-612.

Halstead, John M., A.E. Luloff, and Thomas H. Stevens. 1992. Protest bidders in contingent valuation. *Northeastern Journal of Agricultural and Resource Economics* 21:160-69.

Hanemann, W. Michael. 1984. Welfare evaluations in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics* 66:332-41.

Hanemann, W. Michael. 1994. Valuing the environment through contingent valuation. *The Journal of Economic Perspectives* 8:19-43.

Hanemann W. Michael, John Loomis, and Barbara Kanninen. 1991. Statistical efficiency of double-bounded dichotomous choice contingent valuation. *American Journal of Agricultural Economics* 73:1255-63.

Hanley, Nick, Douglas MacMillan, Robert E. Wright, Craig Bullock, Ian Simpson, Dave Parsisson, and Bob Crabtree. 1998. Contingent valuation versus choice experiments: Estimating the benefits of environmentally sensitive areas in Scotland. *Journal of Agricultural Economics* 49:1-15.

Hanley, Nick, Susana Mourato, and Robert E. Wright. 2001. Choice modelling approaches: A superior alternative for environmental valuation? *Journal of Economic Surveys* 15:435-62.

Hanley, Nick, Felix Schläpfer, and James Spurgeon. 2003. Aggregating the benefits of environmental improvements: Distance-decay functions for use and non-use values. *Journal of Environmental Management* 68:297-304.

Hanley, Nick, Wiktor Adamowicz, and Robert E. Wright. 2005. Price vector effects in choice experiments: An empirical test. *Resource and Energy Economics* 27:227-34.

Hanley, Nick, Bengt Kriström, and Jason F. Shogren. 2009. Coherent arbitrariness: On value uncertainty for environmental goods. *Land Economics* 85:41-50.

Harrison, Glenn W. 2010. Making choice studies incentive compatible. In *Valuing environmental amenities using stated choice studies: A common sense approach to theory and practice,* ed. B.J. Kanninen, 67-110. Dordrecht: Springer.

Harrison, Mark, Dan Rigby, Caroline Vass, Terry Flynn, Jordan Louviere, Katherine Payne. 2014. Risk as an attribute in discrete choice experiments: A systematic review of the literature. *Patient - Patient-Centered Outcomes Research* 7:151–170.

Hausman, Jerry A., ed. 1993. *Contingent valuation: A critical assessment.* Bingley, UK: Emerald Group Publishing Limited.

Hausman, Jerry. 2012. Contingent valuation: From dubious to hopeless. *The Journal of Economic Perspectives* 26:43-56.

Heberlein, Thomas A., Matthew A. Wilson, Richard C. Bishop, and Nora Cate Schaeffer. 2005. Rethinking the scope test as a criterion for validity in contingent valuation. *Journal of Environmental Economic and Management* 50:1-22.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153-61.

Hensher, David A. 2006a. How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics* 21:861-78.

Hensher, David B. 2006b. Revealing differences in willingness to pay due to the dimensionality of stated choice designs: An initial assessment. *Environmental and Resource Economics* 34:7-44.

Hensher, David A., and William H. Greene. 2003. The mixed logit model: The state of practice. *Transportation* 30:133-76.

Hensher, David A., and David Layton. 2010. Parameter transfer of common-metric attributes in choice analysis: Implications for willingness to pay. *Transportation* 37:473-90.

Hensher, David A., and John M. Rose. 2009. Simplifying choice through attribute preservation or non-attendance: Implications for willingness to pay. *Transportation Research Part E: Logistics and Transportation Review* 45:583-90.

Hensher, David A., John M. Rose, and William H. Greene. 2012. Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation* 39:235-45.

Herriges, Joseph A., and Jason F. Shogren. 1996. Starting point bias in dichotomous choice valuation with follow-up questioning. *Journal of Environmental Economics and Management* 30:112-31.

Herriges, Joseph, Catherine Kling, Chi-Chen Liu, and Justin Tobias. 2010. What are the consequences of consequentiality? *Journal of Environmental Economics and Management* 59:67-81.

Hess, Stephane, and Nesha Beharry-Borg. 2012. Accounting for latent attitudes in willingness-to-pay studies: The case of coastal water quality improvements in Tobago. *Environmental and Resource Economics* 52:109-31.

Hess, Stephane, and Andrew Daly, eds. 2010. *Choice modelling: The state of the art and the state of practice*. Proceedings from the Inaugural International Choice Modelling Practice, 30 March-1 April 2009. Cheltenham, UK: Edward Elgar.

Hess, Stephane, and John M. Rose. 2012. Can scale and coefficient heterogeneity be separated in random coefficient models? *Transportation* 39:1225-39.

Ho, Teck H., Noah Lim, and Colin F. Camerer. 2006. Modelling the psychology of consumer and firm behavior with behavioral economics. *Journal of Marketing Research* 43:307-31.

Hoehn, John P., and Alan Randall. 2002. The effect of resource quality information on resource injury perceptions and contingent values. *Resource and Energy Economics* 24:13-31.

Hoehn, John P., Frank Lupi, and Michael D. Kaplowitz. 2010. Stated choice experiments with complex ecosystem changes: The effect of information formats on estimated variances and choice parameters. *Journal of Agricultural and Resource Economics* 35:568-90.

Hole, Arne Risa. 2007. A comparison of approaches to estimating confidence intervals for willingness to pay measures. *Health Economics* 16:827-40.

Hole, Arne Risa. 2011. A discrete choice model with endogenous attribute attendance. *Economics Letters* 110:203-5.

Hole, Arne Risa, Julie Riise Kolstad, and Dorte Gyrd-Hansen. 2013. Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors' prescription behaviour. *Journal of Economic Behavior & Organization* 96:21-31.

Holmes, Thomas P., and Kevin J. Boyle. 2005. Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions. *Land Economics* 81:114-26.

Holmes, Thomas P., and Randall A. Kramer. 1995. An independent sample test of yea-saying and starting point bias in dichotomous-choice contingent valuation. *Journal of Environmental Economics and Management* 29:121-32.

Horne, Pamela, Peter C. Boxall, and Wiktor L. Adamowicz**.** 2005. Multiple-use management of forest recreation sites: A spatially explicit choice experiment. *Forest Ecology and Management* 207:189-99.

Horowitz, John, Kenneth McConnell, and James Murphy. 2013. Behavioral foundations of environmental economics and valuation. In *Handbook on experimental economics and the environment,* eds. John A. List and Michael K. Price, 115-56. Cheltenham, UK: Edward Elgar.

Hoyos, David. 2010. The state of the art of environmental valuation with discrete choice experiments. *Ecological Economics* 69:1595-603.

Huber, Joel, and Kenneth Train. 2001. On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters* 12:259-69.

Hurwicz, Leonid. 1986. Incentive aspects of decentralization. In *Handbook of mathematical* economics, eds. Kenneth J. Arrow and Michael D. Intriligator, Volume 3, 1441-82. Amsterdam: North Holland.

Islam, Towhidul, Jordan J. Louviere, and Paul F. Burke. 2007. Modeling the effects of including/excluding attributes in choice experiments on systematic and random components. *International Journal of Research in Marketing* 24:289-300.

Ivehammar, Pernilla. 2009. The payment vehicle used in CV studies of environmental goods does matter. *Journal of Agricultural and Resource Economics* 34:450-63.

Jacquemet, Nicolas, Alexander James, Stephane Luchini, and Jason F. Shogren. 2016. Referenda under Oath. *Environmental and Resource Economics.* Online First, DOI: 10.1007/s10640-016-0023-5.

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchini, and Jason F. Shogren. 2013. Preference elicitation under oath. *Journal of Environmental Economics and Management* 65:110-32.

Jamison, Julian, Dean Karlan, and Laura Schechter. 2008. To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization* 68:477-88.

Jin, Jianjun, Zhishi Wang, and Shenghong Ran. 2006. Comparison of contingent valuation and choice experiment in solid waste management programs in Macao. *Ecological Economics* 57:430-41.

Johnson, F. Reed, and Kristy E. Mathews. 2001. Sources and effects of utility-theoretic inconsistency in stated-preference surveys. *American Journal of Agricultural Economics* 83:1328-33.

Johnson, F. Reed, Richard W. Dunford, William H. Desvousges, and Melissa Ruby Banzhaf. 2001. Role of knowledge in assessing nonuse values for natural resource damages. *Growth and Change* 32:43-68.

Johnson, F. Reed, Barbara Kanninen, Matthew Bingham and Semra Özdemir. 2006. Experimental design for stated choice studies. In *Valuing environmental amenities using stated choice studies,* ed. Barbara J. Kanninen. New York: Springer.

Johnson, F. Reed, Emily Lancsar, Deborah Marshall, Vikram Kilambi, Axel Mühlbacher, Dean A. Regier, Brian W. Bresnahan, Barbara Kanninen, and John F.P. Bridges. 2013. Constructing experimental designs for discrete choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health* 16:3-13.

Johnston, Robert J. 2006. Is hypothetical bias universal? Validating contingent valuation response using a binding public referendum. *Journal of Environmental Economics and Management* 52:469-81.

Johnston, Robert J., and John C. Bergstrom. 2011. Valuing farmland production: Do empirical results and policy guidance depend on the econometric fine print? *Applied Economic Perspectives and Policy* 33:639-60.

Johnston, Robert J., and Joshua M. Duke. 2007. Willingness to pay for agricultural land preservation and policy process attributes: Does the method matter? *American Journal of Agricultural Economics* 89:1098–115.

Johnston, Robert J., and Mahesh Ramachandran. 2014. Modeling spatial patchiness and hot spots in stated preference willingness to pay. *Environmental and Resource Economics* 59:363-87.

Johnston, Robert J., Thomas F. Weaver, Lynn A. Smith, and Stephen K. Swallow. 1995. Contingent valuation focus groups: Insights from ethnographic interview techniques. *Agricultural and Resource Economics Review* 24:56-69.

Johnston, Robert J., Stephen K. Swallow, and Thomas F. Weaver. 1999. Estimating willingness to pay and resource tradeoffs with different payment mechanisms: An evaluation of a funding guarantee for watershed management. *Journal of Environmental Economics and Management* 38:97-120.

Johnston, Robert J., Stephen A. Swallow, and Dana Marie Bauer. 2002. Spatial factors and stated preference values for public goods: Considerations for rural land use. *Land Economics* 78:481-500.

Johnston, Robert J., Eric T. Schultz, Kathleen Segerson, Elena Y. Besedin, and Mahesh Ramachandran. 2012. Enhancing the content validity of stated preference valuation: The structure and function of ecological indicators. *Land Economics* 88:102-20.

Johnston, Robert J., Eric T. Schultz, Kathleen Segerson, Elena Y. Besedin, and Mahesh Ramachandran. 2016. Biophysical causality and environmental preference elicitation: Evaluating the validity of welfare analysis over intermediate outcomes. *American Journal of Agricultural Economics*, advance access online, doi: 10.1093/ajae/aaw073.

Johnston, Robert J., Daniel Jarvis, Kristy Wallmo, and Daniel K. Lew. 2015. Multiscale spatial pattern in nonuse willingness to pay: Applications to threatened and endangered marine species. *Land Economics* 91:739-61.

Jorgensen, Bradley S. 1999. Focus groups in the contingent valuation process: A real contribution or missed opportunity. *Journal of Economic Psychology* 20:485-9.

Jorgensen, Bradley S., and Geoffrey J. Syme. 2000. Protest responses and willingness to pay: Attitude toward paying for stormwater pollution abatement. *Ecological Economics* 33:251-65.

Jorgensen, Bradley S., Geoffrey J. Syme, Brian J. Bishop, and Blair E. Nancarrow. 1999. Protest responses in contingent valuation. *Environmental and Resource Economics* 14:131-50.

Jørgensen, Sisse Liv, Søren Bøye Olsen, Jacob Ladenburg, Louise Martinsen,Stig Roar Svenningsen, and Berit Hasler. 2013. Spatially induced disparities in users' and non-users' WTP for water quality improvements—Testing the effect of multiple substitutes and distance decay. *Ecological Economics* 92:58–66.

Kahneman, Daniel. 2003. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review* 93:1449-75.

Kahneman, Daniel, and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47:263-92.

Kahneman Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgement under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kanninen, Barbara J. 1993a. Design of sequential experiments for contingent valuation studies. *Journal of Environmental Economics and Management* 25:S1-S11.

———. 1993b. Optimal experimental design for double-bounded dichotomous choice contingent valuation. *Land Economics* 69:138-46.

———. 1995. Bias in discrete response contingent valuation. *Journal of Environmental Economics and Management* 28:114-25.

Kanninen, Barbara J., ed. 2010. *Valuing environmental amenities using stated choice studies: A common sense approach to theory and practice*. Dordrecht, Netherlands: Springer.

Kaplowitz, Michael D., and John P. Hoehn. 2001. Do focus groups and individual interviews reveal the same information for natural resource valuation? *Ecological Economics* 36:237-47.

Kaplowitz, Michael D., Frank Lupi, and John P. Hoehn. 2004. Multiple methods for developing and evaluating a stated-choice questionnaire to value wetlands. In *Methods for testing and evaluating survey questionnaires*, eds. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. Hoboken, NJ: Wiley.

Kataria, Mitesh, Ian Bateman, Tove Christensen, Alex Dubgaard, Berit Hasler, Stephanie Hime, Jacob Ladenburg, Gregor Levin, Louise Martinsen, and Carsten Nissen. 2012. Scenario realism and welfare estimates in choice experiments—A non-market valuation study on the European water framework directive. *Journal of Environmental Management* 94:25-33.

Kaul, Sapna, Kevin J. Boyle, Nicolai V. Kuminoff, Christopher E. Parmeter, and Jaren C. Pope. 2013. What can we learn from benefit transfer errors? Evidence from 20 years of research on convergent validity. *Journal of Environmental Economics and Management* 66:90-104.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly* 64:125-48.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly* 70:759-79.

Kerr, Geoffrey N., and Basil M.H. Sharp. 2006. Transferring mitigation values for small streams. *In Choice modelling and the transfer of environmental values,* eds. John Rolfe and Jeff Bennett, 136-63. Cheltenham, UK: Edward Elgar.

Kerr, Geffrey N., and Basil M.H. Sharp. 2010. Choice experiment adaptive design benefits: A case study. *Australian Journal of Agricultural and Resource Economics* 54:407-20.

Kessels, Roselinde, Peter Goos, and Martina Vandebroek. 2006. A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research* 43:409-19.

Kessels, Roselinde, Bradley Jones, Peter Goos, and Martina Vandebroek. 2009. An efficient algorithm for constructing Bayesian optimal choice designs. *Journal of Business & Economic Statistics* 27: 279-91.

Kim, Younjun, Catherine L. Kling, and Jinhua Zhao. 2015. Understanding behavioral explanations of the WTP-WTA divergence through a neoclassical lens: Implications for environmental policy. *Annual Review of Resource Economics* 7:169-87.

Kjaer, Trine, Mickael Bech, Dorte Gyrd-Hansen, and Kristian Hart-Hansen. 2006. Ordering effect and price sensitivity in discrete choice experiments: Need we worry? *Health Economics* 15:1217-28.

Kling, Catherine L., Daniel J. Phaneuf, and Jinhua Zhao. 2012. From Exxon to BP: Has some number become better than no number? *The Journal of Economic Perspectives* 26:3-26.

Knetsch, Jack L. 2007. Biased valuations, damage assessments, and policy choices: The choice of measure matters. *Ecological Economics* 63: 684-89.

Kriström, Bengt. 1997. Spike models in contingent valuation. *American Journal of Agricultural Economics* 79:1013-23.

Krosnick, Jon A., and Stanley Presser. 2010. Question and questionnaire design. *Handbook of Survey Research* 2:263-314.

Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, and V. Kerry Smith, et al. 2002. The impact of “no opinion” response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly* 66:371-403.

Krueger, Richard A., and Mary Anne Casey. 2015. *Focus groups: A practical guide for applied research*, 5th ed. Thousand Oaks, CA: Sage Publications.

Krupnick, Alan, and Wiktor L. (Vic) Adamowicz. 2006. Supporting questions in stated-choice studies. In *Valuing environmental amenities using stated choice studies,* ed. Barbara J. Kanninen. New York: Springer.

Krupnick, Alan, and David A. Evans. 2008. Sample representativeness: Implications for administering and testing stated preference surveys. In *Proceedings of a workshop held at Resources for the Future,* sponsored by the National Center for Environmental Economics and the US Environmental Protection Agency, October 2, 2006, Appendix D. Washington, DC: Resources for the Future.

Kuhfeld, Warren F. 2005. Marketing research methods in SAS. *Experimental design, choice. conjoint, and graphical techniques*. Cary, NC: SAS Institute TS-722.

Lancaster, Kelvin J. 1966. A new approach to consumer theory. *The Journal of Political Economy* 74:132-57.

Lancsar, Emily, and Jordan J. Louviere. 2006. Deleting "irrational" responses from discrete choice experiments: A case of investigating or imposing preferences? *Health Economics* 15:797-811.

Lancsar, Emily, and Jordan J. Louviere. 2008. Conducting discrete choice experiments to inform healthcare decision making: A user's guide. *PharmacoEconomics* 26:661-77.

Lareau, T., and D. Rae. 1989. Valuing willingness to pay for diesel odor reductions: An application of contingent ranking technique. *Southern Economic Journal* 55:728-42.

LaRiviere, Jacob, Mikolaj Czajkowski, Nick Hanley, Margrethe Aanesen, Jannike Falk-Petersen, and Dugald Tinch. 2014. The value of familiarity: Effects of knowledge and objective signals on willingness to pay for public goods. *Journal of Environmental Economics and Management* 68:376-89.

Layton, David F., and S. Todd Lee. 2006. Embracing model uncertainty: Strategies for response pooling and model averaging. *Environmental and Resource Economics* 34:51-85.

Leamer, Edward E. 1983. Let's take the con out of econometrics. *The American Economic Review* 73:31-43.

Lee, Jaeseung Jason, and Trudy Ann Cameron. 2008. Popular support for climate mitigation: Evidence from a general population mail survey. *Environmental and Resource Economics* 41:223-48.

Leggett, Christopher G., Naomi S. Kleckner, Kevin J. Boyle, John W. Dufield, and Robert Cameron Mitchell. 2003. Social desirability bias in contingent valuation studies administered through in-person interviews. *Land Economics* 79:561-75.

Lewbel, Arthur. 2000. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics* 97:145-77.

Lewbel, Arthur, Daniel McFadden, and Oliver Linton. 2011. Estimating features of a distribution from binomial data. *Journal of Econometrics* 162:170-88.

Lindhjem, Henrik, and Ståle Navrud. 2011a. Are internet surveys an alternative to face-to-face interviews in contingent valuation? *Ecological Economics* 70:1628-37.

———. 2011b. Using internet in stated preference surveys: A review and comparison of survey modes. *International Review of Environmental and Resource Economics* 5:309-51.

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008. A comparison of address-based sampling (ABS) versus random digit-dialing (RDD) for general population surveys. *Public Opinion Quarterly* 72:6-27.

Lipkus, Isaac M., and Justin G. Hollands. 1999. The visual communication of risk. *Journal of the National Cancer Institute. Monographs* 25:149-63.

List, John A. 2009. The IRB is key in field experiments. *Science* 323 (5915): 713-14.

List, John A. 2011. Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives* 25:3-15.

List, John A., and Craig A. Gallet. 2001. What experimental protocol influence disparities between actual and hypothetical stated values? Evidence from a meta-analysis. *Environmental and Resource Economics* 20:241-54.

Little, Joseph, and Robert Berrens. 2004. Explaining disparities between actual and hypothetical stated values: Further investigation using meta-analysis. *Economics Bulletin* 3:1-13.

Lohr, Sharon L., and J. Michael Brick. 2014. Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology* 2:388-409.

Loomis, John B. 1989. Test-retest reliability of the contingent valuation method: A comparison of general population and visitor responses. *American Journal of Agricultural Economics* 71:76-84.

———. 1996. How large is the extent of the market for public goods: Evidence from a nationwide contingent valuation survey. *Applied Economics* 28:779-82.

———. 2000.Vertically summing public good demand curves: An empirical comparison of economic versus political jurisdiction. *Land Economics* 76:312-21.

———. 2014. Strategies for overcoming hypothetical bias in stated preference surveys. *Journal of Agricultural and Resource Economics* 39:34-46.

Loomis John B., and Pierre H. duVair. 1993. Evaluating the effect of alternative risk communication devices on willingness to pay: Results from a dichotomous choice contingent valuation experiment. *Land Economics* 69:287-98.

Loomis, John B., and Randall S. Rosenberger. 2006. Reducing barriers in future benefit transfers: Needed improvements in primary study design and reporting. *Ecological Economics* 60:343-50.

Loomis, John, Kerri Traynor, and Thomas Brown. 1999. Trichotomous choice: A possible solution to dual response objectives in dichotomous choice contingent valuation questions. *Journal of Agricultural and Resource Economics* 24:572-83.

Louviere, Jordan J., Terry N. Flynn, and Richard T. Carson. 2010. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* 3:57-72.

Louviere, Jordan J., Towhidul Islam, Nada Wasi, Deborah Street, and Leonie Burgess. 2008. Designing discrete choice experiments: Do optimal designs come at a price? *Journal of Consumer Research* 35: 360-75.

Louviere, Jordan J., David A. Hensher, and Joffre D. Swait. 2000. *Stated choice methods: Analysis and applications.* New York: Cambridge University Press.

Lovell, Michael C. 1983. Data Mining. *The Review of Economics and Statistics* 65:1-12.

Luce, R. Duncan, and John W. Tukey 1964. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* 1:1-27.

Lundhede, Thomas, Jette Bredahl Jacobsen, Nick Hanley, Niels Strange, and Bo Jellesmark Thorsen. 2015. Incorporating outcome uncertainty and prior outcome beliefs in stated preference. *Land Economics* 91:296-316.

Lundquist, Peter and Carl-Erik Särndal. 2013. Aspects of responsive design with applications to the Swedish living conditions survey. *Journal of Official Statistics* 29:557-82.

Lunt, Peter. 1999. Comments on Chilton and Hutchinson: Beyond measurement issues in the focus group method. *Journal of Economic Psychology* 20:491-4.

Lusk, Jayson L. 2003. Effects of cheap talk on consumer willingness-to-pay for golden rice. *American Journal of Agricultural Economics* 85:840-56.

Lusk, Jayson L., and Ted C. Schroeder. 2004. Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics* 86:467-82.

Maddala, G.S. 1986. *Limited-dependent and qualitative variables in econometrics.* New York: Cambridge University Press.

Madureira, Lívia, Luis C. Nunes, and José M.L. Santos. 2005. Valuing multi-attribute environmental changes: Contingent valuation and choice experiments. Paper presented at 14th Annual Conference of the European Association of Environmental and Resource Economists, Bremen, June 23-26, 2005.

Macmillan Douglas C., Lorna Philip, Nick Hanley, and Begona Alvarez-Farizo. 2002. Valuing the non-market benefits of wild goose conservation: A comparison of interview and group-based approaches. *Ecological Economics* 43:49-59.

Manski, Charles F. 1977. The structure of random utility models. *Theory and Decision* 8:229-54.

Manski, Charles F. 2008. *Identification for prediction and decision.* Cambridge, MA: Harvard University Press.

Mariel, Petr, Jürgen Meyerhoff, and Stephane Hess. 2015. Heterogeneous preferences toward landscape externalities of wind turbines: Combining choices and attitudes in a hybrid model. *Renewable and Sustainable Energy Reviews* 41:647-57.

Marley, Anthony A.J., and Jordan J. Louviere. 2005. Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology* 49:464-80.

Marsden, Peter V., and James D. Wright. 2010. *Handbook of Survey Research* (2nd edition). Bingley, UK: Emerald Group Publishing Limited.

Marwell, Gerald, and Ruth E. Ames. 1980. Experiments on the provision of public goods. II. Provision points, stakes, experience, and the free-rider problem. *American Journal of Sociology* 85:926-37.

Mazzotta, Marisa J., and James J. Opaluch. 1995. Decision making when choices are complex: A test of Heiner's hypothesis. *Land Economics* 71:500-15.

McConnell, Kenneth E. 1990. Models for referendum data: The structure of discrete choice models for contingent valuation. *Journal of Environmental Economics and Management* 18:19-34.

McDaniel, Tanga, and Chris Starmer. 1998. Experimental economics and deception: A comment. *Journal of Economic Psychology* 19:403-9.

McFadden, Daniel. 1986. The choice theory approach to market research. *Marketing Science* 5:275-297.

McFadden, Daniel. 2014. The New Science of Pleasure: Consumer Behavior and the Measurement of Well-Being. In *Handbook of choice modelling,* eds. Stephane Hess and Andew Daly, 7-48. Cheltenham, UK: Edward Elgar.

McFadden, Daniel, and Kenneth Train. 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15:447-70.

Meijer, Erik, and Jan Rouwendal. 2006. Measuring welfare effects in models with random coefficients. *Journal of Applied Econometrics* 21:227-44.

Mercer, Andrew, Andrew Caporaso, David Cantor, and Reanne Townsend. 2015. How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly* 79:105-29.

Merkle, Daniel, and Murray Edelman. 2002. Nonresponse in exit polls: A comprehensive analysis. In *Survey nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J.A. Little, 243-58. New York: Wiley.

Messer, Benjamin L., and Don A. Dillman. 2011. Surveying the general public over the internet using address-based sampling and mail contact features. *Public Opinion Quarterly* 75:429-57.

Messonnier, Mark L., John C. Bergstrom, Christopher M. Cornwell, R. Jeff Teasley, and H. Ken Cordell. 2000. Survey response-related biases in contingent valuation: Concepts, remedies, and empirical application to valuing aquatic plant management. *American Journal of Agricultural Economics* 82:438-50.

Meyer, Andrew. 2013. Intertemporal valuation of river restoration. *Environmental and Resource Economics* 54:41-61.

Meyerhoff, Jürgen, and Ulf Liebe. 2006. Protest beliefs in contingent valuation: Explaining their motivation. *Ecological Economics* 57:583-94.

———. 2008. Do protest responses to a contingent valuation question and a choice experiment differ? *Environmental and Resource Economics* 39:433-46.

———. 2009. Status quo effect in choice experiments: Empirical evidence on attitudes and choice task complexity. *Land Economics* 85:515-28.

Meyerhoff, Jürgen, Anna Bartczak, and Ulf Liebe. 2012. Protester or non-protestor: A binary state? On the use (and non-use) of of latent class models to analyse protesting in economic valuation. *Australian Journal of Agricultural and Resource Economics* 56:438-54.

Meyerhoff, Jürgen, Malte Oehlmann, and Priska Weller. 2015. The influence of design dimensions on stated choices in an environmental context. *Environmental and Resource Economics* 61:385-407.

Millar, Morgan M., and Don A. Dillman. 2011. Improving response to web and mixed-mode surveys. *Public Opinion Quarterly* 75:249-69.

Mitchell, Robert Cameron, and Richard T. Carson. 1989. *Using surveys to value public goods: The contingent valuation method.* Washington, DC: Resources for the Future.

Mogas, Joan, Pere Riera, and Jeff Bennett. 2006. A comparison of contingent valuation and choice modelling with second-order interactions. *Journal of Forest Economics* 12:5-30.

Morey, Edward R., Vijaya R. Sharma, and Anders Karlstrom. 2003. A simple method of incorporating income effects into logit and nested-logit models: Theory and application. *American Journal of Agricultural Economics* 85:248-53.

Morgan, David L. 1997. *Focus groups as qualitative research*. Qualitative Research Methods Series, Volume 16, 2nd edition. Thousand Oaks, CA: Sage Publications.

Mørkbak, Morten Raun, and Søren Bøye Olsen. 2015. A within-sample investigation of test-retest reliability in choice experiment surveys with real economic incentives. *Australian Journal of Agricultural and Resource Economics* 59:375-92.

Morrison, Mark. 2000. Aggregation biases in stated preference studies. *Australian Economics Papers* 39:215-30.

Morrison, Mark. 2002. Rethinking contingent valuation: Ethics versus defensibility? *International Journal of Agricultural Resources, Governance and Ecology* 2:22-36.

Morrison, Mark D., Blamey, Russell K., Jeff W. Bennett. 2000. Minimising payment vehicle bias in contingent valuation studies. *Environmental and Resource Economics* 16:407-422.

Moser, Riccarda, Roberta Raffaelli, and Sandra Notaro. 2013. Testing hypothetical bias with a real choice experiment using respondents' own money. *European Review of Agricultural Economics* 41:25-46.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. 2005a. A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics* 30:313-25.

Murphy, James J., Thomas H. Stevens, and Darryl Weatherhead. 2005b. Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism? *Environmental and Resource Economics* 30:327-43.

National Research Council. 2013. Nonresponse in social science surveys: A research agenda. Roger Tourangeau and Thomas J. Plewes, eds. Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olsen, Søren Bøye, and Jürgen Meyerhoff. 2016. Will the alphabet soup of design criteria affect discrete choice experiment results? *European Review of Agricultural Economics* doi:10.1093/erae/jbw014.

Pate, Jennifer, and John Loomis. 1997. The effect of distance on willingness to pay values: A case study of wetlands and salmon in California. *Ecological Economics* 20:199-207.

Pearce, David, and Ece Özdemiroglu, et al. 2002. *Economic valuation with stated preference techniques: Summary guide*. London: Department for Transport, Local Government and the Regions.

Petrin, Amil, and Kenneth Train. 2003. *Omitted product attributes in discrete choice models.* Cambridge, MA: National Bureau of Economic Research.

Peytchev, Andy. 2009. Survey breakoff. *Public Opinion Quarterly* 73:74-97.

Pitchforth Emma, Verity Watson, Janet Tucker, Mandy Ryan, Edwin van Teijlingen, Jane Farmer, Jillian Ireland , Elizabeth Thomson, Alice Kiger, and Helen Bryers. 2008. Models of intrapartum care and women's trade-offs in remote and rural Scotland: A mixed-methods study. *British Journal of Obstetrics and Gynaecology* 115:560-69.

Poe, Gregory L., Jeremy E. Clark, Daniel Rondeau, and William D. Schulze. 2002. Provision point mechanisms and field validity tests of contingent valuation. *Environmental and Resource Economics* 23:105-31.

Powe, Neil A. 2007. *Redesigning environmental valuation. Mixing methods with stated preference techniques.* Cheltenham, UK: Edward Elgar.

Powe, Neil A., and Ian H. Bateman. 2004. Investigating insensitivity to scope: A split-sample test of perceived scheme realism. *Land Economics* 80:258-71.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. 2004. Methods for testing and evaluating survey questions. *Public Opinion Quarterly* 68:109-30.

Promberger, Marianne, Paul Dolan, and Theresa M. Marteau. 2012. "Pay them if it works": Discrete choice experiments on the acceptability of financial incentives to change health related behaviour. *Social Science and Medicine* 75:2509-14.

Randall, Alan. 1994. A difficulty with the travel cost method. *Land Economics* 70:88-96.

Ready, Richard C., Jean C. Buzby, and Dayuan Hu. 1996. Differences between continuous and discrete contingent valuation estimates. *Land Economics* 72:397-411.

Richard C. Ready, Patricia A. Champ, and Jennifer L. Lawton. 2010. Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Economics* 86:363-81.

Ready, Richard C., John C. Whitehead, and Glenn C. Blomquist. 1995. Contingent valuation when respondents are ambivalent. *Journal of Environmental Economics and Management* 29:181-96.

Revelt, David, and Kenneth Train. 1998. Mixed logit with repeated choices. *The Review of Economics and Statistics* 80: 647-57.

Riach, Peter A., and Judith Rich. 2004. Deceptive field experiments of discrimination: Are they ethical? *Kyklos* 57:457-70.

Riera, Pere, Giovanni Signorello, Mara Thiene, Pierre-Alexandre Mahieu, Ståle Navrud , Pamela Kaval, Benedicte Rulleau, Robert Mavsar, Lívia Madureira, Jürgen Meyerhoff, Peter Elsasser, Sandra Notaro, Maria De Salvo, Marek Giergiczny, and Simona Dragoi. 2012. Non-market valuation of forest goods and services: Good practice guidelines. *Journal of Forest Economics* 18:259-70.

Roberts, David C., Tracy Boyer, and Jayson L. Lusk. 2008. Preferences for environmental quality under uncertainty. *Ecological Economics* 66:584-93.

Rolfe, John, and Jeff Bennett. 2009. The impact of offering two versus three alternatives in choice modelling experiments. *Ecological Economics* 68:1140-8.

Rolfe, John, Jeff Bennett, and Jordan Louviere. 2002. Stated values and reminders of substitute goods: Testing for framing effects with choice modelling. *Australian Journal of Agricultural and Resource Economics* 46:1-20.

Rolfe, John, and Jill Windle. 2012. Distance decay functions for iconic assets: Assessing national values to protect the health of the Great Barrier Reef in Australia. *Environmental and Resource Economics* 53:347-65.

Rolfe, John, and Jill Windle. 2015. Do respondents adjust their expected utility in the presence of an outcome certainty attribute in a choice experiment? *Environmental and Resource Economics* 60:125-42.

Rollins, Kimberly, and Audrey Lyke. 1998. The case for diminishing marginal existence values. *Journal of Environmental Economics and Management* 36:324-44.

Rollins, Kimberly, MDR Evans, Mimako Kobayashi, and Anita Castledine. 2010. *Willingness to pay estimation when protest beliefs are not separable from the public good definition.* Reno, NV: University of Nevada, Department of Resource Economics, UNR Joint Economics Working Paper Series, Working Paper No. 10-002.

Rondeau, Daniel, William D. Schulze, and Gregory L. Poe. 1999. Voluntary revelation of the demand for public goods using a provision point mechanism. *Journal of Public Economics* 72:455-70.

Rose, John M., Michiel C.J. Bliemer, David A. Hensher, and Andrew T. Collins. 2008. Designing efficient stated choice experiments in the presence of reference alternatives. *Transportation Research Part B: Methodological* 42: 395-406.

Rose, John M., Riccardo Scarpa, and Michiel C.J. Bliemer. 2009. Incorporating model uncertainty into the generation of efficient stated choice experiments: A model averaging approach. Working Paper ITLS-WP-09-08. Sydney: Institute of Transport and Logistics Studies, University of Sydney.

Rose, John M., and Michiel C.J. Bliemer. 2009. Constructing efficient stated choice experimental designs. *Transport Reviews* 29:587-617.

Rose, John, and Michiel C.J. Bliemer. 2014. Stated choice experimental design theory: The who, the what and the why. In *Handbook of choice modelling*, eds. Stephane Hess, and Andrew Daly, chapter 7. Cheltenham, UK: Edward Elgar.

Rose, Steven K., Jeremy Clark, Gregory L. Poe, Daniel Rondeau, and William D. Schulze. 2002. The private provision of public goods: Tests of a provision point mechanism for funding green power programs. *Resource and Energy Economics* 24:131-55.

Rosenberger, Randall S., and Robert J. Johnston. 2009. Selection effects in meta-analysis and benefit transfer: Avoiding unintended consequences. *Land Economics* 85:410-28.

Rossi, Peter H., James D. Wright, and Andy B. Anderson. eds. 2013. *Handbook of survey research.* Cambridge, MA: Academic Press.

Rousu, Matthew C., Gregory Colson, Jay R. Corrigan, Carola Grebitus, and Maria L. Loureiro. 2015. Deception in experiments: Towards guidelines on use in applied economics research. *Applied Economics Perspectives and Policy* 37:524-36.

Rowe, Robert D., William D. Schulze, and William S. Breffle. 1996. A test for payment card biases. *Journal of Environmental Economics and Management* 31:178-85.

Ryan, Mandy. 1999. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Social Science and Medicine* 48:535-46.

Ryan, Mandy. 2004. A comparison of stated preference methods for estimating monetary values. *Health Economics* 13:291-6.

Ryan, Mandy, and Angela Bate. 2001. Testing the assumptions of rationality, continuity and symmetry when applying discrete choice experiments in health care. *Applied Economics Letters* 8:59-63.

Ryan, Mandy, and Karen Gerard. 2003. Using discrete choice experiments to value health care programmes: Current practice and future research reflections. *Applied Health Economics and Policy Analysis* 2:55-64.

Ryan, Mandy, and Verity Watson. 2009. Comparing welfare estimates from payment card contingent valuation and discrete choice experiments. *Health Economics* 18:389-401.

Ryan, Mandy, Karen Gerard, and Mabel Amaya-Amaya, eds. 2008. *Using discrete choice experiments to value health and health care.* Vol. 11 in The Economics of Non-Market Goods and Resources. Dordrecht, Netherlands: Springer.

Ryan, Mandy, and Fernando San Miguel. 2003. Revisiting the axiom of completeness in health care. *Health Economics* 12:295-307.

Ryan, Mandy, Emmanouil Mentzakis, Sutthi Jareinpituk, and John Cairns. Forthcoming. Testing the external validity of contingent valuation: Comparing hypothetical and real payments. *Health Economics*.

Ryan, Mandy, Verity Watson, and Vikki Entwistle. 2009. Rationalising the 'irrational': A think aloud study of discrete choice experiment responses. *Health Economics* 18:321-36.

Sanchirico, James N., Daniel K. Lew, Alan C. Haynie, David M. Kling, and David F. Layton. 2013. Conservation values in marine ecosystem-based management. *Marine Policy* 38:523-30.

Sándor, Zsolt, and Michael Wedel. 2002. Profile construction in experimental choice designs for mixed logit models. *Marketing Science* 21:455-75.

Sandorf, Erlend Dancke, Danny Campbell, and Nick Hanley. 2016. Disentangling the influence of knowledge on attribute non-attendance. *Journal of Choice Modelling* doi: 10.1016/j.jocm.2016.09.003.

Särndal, Carl-Erik. 2011. The 2010 Morris Hansen Lecture dealing with survey response in data collection, in estimation. *Journal of Official Statistics* 27:1-21.

Scarpa, Riccardo, and Anna Alberini, eds. 2005. *Applications of simultaneous methods in environmental and resource economics.* Dordrecht: Springer.

Scarpa, Riccardo, and Ian Bateman. 2000. Efficiency gains afforded by improved bid design versus follow-up valuation questions in discrete-choice CV studies. *Land Economics* 76:299-311.

Scarpa, Riccardo, and John M. Rose. 2008. Design efficiency for non-market valuation with choice modelling: How to measure it, what to report and why. *Australian Journal of Agricultural and Resource Economics* 52:253-82.

Scarpa, Riccardo, and Mara Thiene. 2005. Destination choice models for rock climbing in the Northeastern Alps: A latent-class approach based on intensity of preferences. *Land Economics* 81:426-44.

Scarpa, Riccardo, Danny Campbell, and W. George Hutchinson. 2007. Benefit estimates for landscape improvements: Sequential Bayesian design and respondents' rationality in a choice experiment study. *Land Economics* 83:617-34.

Scarpa, Riccardo, Mara Thiene, and David A. Hensher. 2010. Monitoring choice task attribute attendance in nonmarket valuation of multiple park management services: Does it matter? *Land Economics* 86:817-39.

Scarpa, Riccardo, Mara Thiene, and Francesco Marangon. 2008a. Using flexible taste distributions to value collective reputation for environmentally friendly production methods. *Canadian Journal of Agricultural Economics* 56:145-62.

Scarpa, Riccardo, Mara Thiene, and Kenneth Train. 2008b. Utility in willingness to pay space: A tool to address confounding random scale effects in destination choice to the Alps. *American Journal of Agricultural Economics* 90:994-1010.

Scarpa, Riccardo, Timothy J. Gilbride, Danny Campbell, and David A. Hensher. 2009. Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics* 36:151-74.

Scarpa, Riccardo, Sandra Notaro, Jordan Louviere, and Roberta Raffaelli. 2011. Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. *American Journal of Agricultural Economics* 93:813-28.

Ščasný, Milan, and Anna Alberini. 2012. Valuation of mortality risk attributable to climate change: Investigating the effect of survey administration modes on a VSL. *International Journal Environmental Research and Public Health* 9:4760-81.

Schaafsma, Marije, Roy Brouwer, and John Rose. 2012. Directional heterogeneity in WTP models for environmental valuation. *Ecological Economics* 79:21-31.

Schaafsma, Marije, Roy Brouwer, Alison Gilbert, Jeroen van den Bergh, and Alfred Wagtendonk. 2013. Estimation of distance-decay functions to account for substitution and spatial heterogeneity in stated preference research. *Land Economics* 89:514-37.

Schkade, David A., and John W. Payne. 1994. How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation. *Journal of Environmental Economics and Management* 26:88-109.

Schoemaker, Paul J.H. 1982. The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature* 20:529-63.

Schouten, Barry, Fanny Cobben, and Jelke Bethlehem. 2009. Indicators for the representativeness of survey response. *Survey Methodology* 35:101-13.

Schultz, Eric T., Robert J. Johnston, Kathleen Segerson, and Elena Y. Besedin. 2012. Integrating ecology and economics for restoration: Using ecological indicators in valuation of ecosystem services. *Restoration Ecology* 20:304-10.

Schuman, Howard, and Stanley Presser. 1979. The assessment of “no opinion” in attitude surveys. *Sociological Methodology* 10:241-75.

Scott, Anthony. 1965. The valuation of game resources: Some theoretical aspects. *Canadian Fisheries Report* 4:27-47.

Shaw, W. Douglass, and Justin Baker. 2010. Models of location choice and willingness to pay to avoid hurricane risks for Hurricane Katrina evacuees. *International Journal of Mass Emergencies and Disasters* 28:87-114.

Sillano, Mauricio, and Juan de Dios Ortúzar. 2005. Willingness-to-pay estimation with mixed logit models: Some new evidence. *Environment and Planning A* 37:525-50.

Singer, Eleanor, and Cong Ye. 2013. The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science* 645:112-41.

Smith, Richard D. 2006. It's not just what you do, it's the way that you do it. The effect of different payment card formats and survey administration on willingness to pay for health gain. *Health Economics* 15:281-93.

Smith, V. Kerry. 2006b. Judging quality. In *Valuing environmental amenities using stated choice studies: A common sense approach to theory and practice*, Barbara J. Kanninen, ed., chapter 11. Dordrecht: Springer.

Smith, V. Kerry, and Laura L. Osborne. 1996. Do contingent valuation estimates pass a "scope" test? A meta-analysis. *Journal of Environmental Economics and Management* 31:287-301.

Stanley, T.D. and Hristos Doucouliagos. 2012. *Meta-regression analysis in economics and business.* New York: Routledge.

Stevens, Thomas H., Maryam Tabatabaei, and Daniel Lass. 2013. Oaths and hypothetical bias. *Journal of Environmental Management* 127:135-41.

Sutherland, Ronald J. and Richard G. Walsh. 1985. Effect of distance on the preservation value of water quality. *Land Economics* 61: 281-91.

Swait, Joffre, and Wiktor Adamowicz. 2001a. Choice environment, market complexity, and consumer behavior: A theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organizational Behavior and Human Decision Processes* 86:141-67.

Swait, Joffre, and Wiktor Adamowicz. 2001b. The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research* 28:135-48.

Swait, Joffre, and Jordan Louviere. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research* 30:305-14.

Swedish Environmental Protection Agency (EPA). 2006. An instrument for assessing the quality of environmental valuation studies. Naturvårdsverket, SE-106 48 Stockholm, Sweden.

Tancreto, Jennifer, Mary Frances Zelenak, Michelle Ruiter, and Brenna Matthews. 2012. *2011 American Community Survey Internet Tests: Results from first test in April 2011*. Washington, DC: US Census Bureau, Decennial Statistics Studies Division, #ACS12-RER-13.

Taylor, Laura O., Mark D. Morrison, and Kevin J. Boyle. 2010. Exchange rules and the incentive compatibility of choice experiments. *Environmental and Resource Economics* 47:197-220.

Teisl, Mario F., Kevin J. Boyle, Daniel W. McCollum, and Stephen D. Reiling. 1995. Test-retest reliability of contingent valuation with independent sample pretest and posttest control groups. *American Journal of Agricultural Economics* 77:613-19.

Thayer, Mark A. 1981. Contingent valuation techniques for assessing environmental impacts: Further evidence. *Journal of Environmental Economics and Management* 8:27-44.

Thurstone, Louis L. 1927. A law of comparative judgment. *Psychological Review* 34:273-86.

Torres, Cati, Nick Hanley, and Antoni Riera. 2011. How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. *Journal of Environmental Economics and Management* 62:111-21.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2004. Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly* 68:368-93.

Tourangeau, Roger, and Ting Yan. 2007. Sensitive questions in surveys. *Psychological Bulletin* 133:859-83.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The science of web surveys.* New York: Oxford University Press.

Train, Kenneth E. 1998. Recreation demand models with taste differences over people. *Land Economics* 74:230-39.

Train, Kenneth E. 2009. *Discrete choice methods with simulation*. New York: Cambridge University Press.

Train, Kenneth and Melvyn Weeks. 2005. Discrete choice models in preference space and willing-to-pay space. In *Applications of simulation methods in environmental and resource economics*, Riccardo Scarpa, and Anna Alberini, eds., 1-16. Dordrecht, Netherlands: Springer.

Train, Kenneth E., Daniel L. McFadden, and Andrew A. Goett. 1987. Consumer attitudes and voluntary rate schedules for public utilities. *The Review of Economics and Statistics* 69:383-91.

US Environmental Protection Agency (EPA). 2014. Guidelines for *preparing economic analyses.* Washington DC: Office of Policy, National Center for Environmental Economics.

Vajjhala, Shalini P., Anna Mische John, and David A. Evans. 2008. Determining the extent of market and extent of resource for stated preference survey design using mapping methods. RFF *Environmental Economics Working Paper Series.* No. 2008-09.

Van den Berg, Vincent A.C., Eric Kroes, and Erik T. Verhoef. 2010. Biases in willingness-to-pay measures from multinomial logit estimates due to unobserved heterogeneity. Amsterdam: Tinbergen Institute Discussion Paper 10-014/3.

Varian, Hal R. 1992. *Microeconomic analysis.* New York: W.W. Norton.

Veall, Michael R. 1992. Bootstrapping the process of model selection: An econometric example. *Journal of Applied Econometrics* 7:93-9.

Veisten, Knut, Hans Fredrik Hoen, Ståle Navrud, and Jon Strand. 2004. Scope insensitivity in contingent valuation of complex environmental amenities. *Journal of Environmental Management* 73:317-31.

Verma, Inder M. 2014. Editorial expression of concern: Experimental evidence of massive scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111:10779.

Vermeulen, Bart, Peter Goos, Riccardo Scarpa, and Martina Vandebroek. 2011. Bayesian conjoint choice designs for measuring willingness to pay. *Environmental and Resource Economics* 48:129-49.

Viscusi, W. Kip. 2014. The value of individual and societal risks to life and health. *Handbook of the Economics of Risk and Uncertainty* 1:385-452.

von Haefen, Roger H., D. Matthew Massey, and Wiktor L. Adamowicz. 2005. Serial nonparticipation in repeated discrete choice models. *American Journal of Agricultural Economics* 87:1061-76.

Vossler, Christian A. 2016. Chamberlin meets Ciriacy-Wantrup: Using insights from experimental economics to inform stated preference research. *Canadian Journal of Agricultural Economics* 64:33-48.

Vossler, Christian A., and J. Scott Holladay. 2016. Alternative value elicitation formats in contingent valuation: A new hope. Working Paper No. 2016-02. Knoxville, TN: University of Tennessee, Department of Economics.

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* 58:338-45.

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* 45:631-49.

Vossler, Christian A., and Sharon B. Watson. 2013. Understanding the consequences of consequentiality: Testing the validity of stated preferences in the field. *Journal of Economic Behavior & Organization* 86:137-47.

Vossler, Christian A., Joe Kerkvliet, Stephen Polasky, and Olesya Gainutdinova. 2003. Externally validating contingent valuation: An open-space survey and referendum in Corvallis, Oregon. *Journal of Economic Behavior & Organization* 51:261-77.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. 2012. Truth in consequentiality: Theory and field evidence on discrete choice experiments. *American Economic Journal: Microeconomics* 4:145-71.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70:129-33.

Watanabe, Masahide. 2010. Nonparametric estimation of mean willingness to pay from discrete response valuation data. *American Journal of Agricultural Economics* 92:1114-35.

Watanabe, Masahide, and Kota Asano. 2009. Distribution free consistent estimation of mean WTP in dichotomous choice contingent valuation. *Environmental and Resource Economics* 44:1-10.

Whitehead, John C. 1991. Environmental interest group behavior and self-selection bias in contingent valuation mail surveys. *Growth and Change* 22:10-20.

Whitehead, John C., Peter A. Groothuis, and Glenn C. Blomquist. 1993. Testing for non-response and sample selection bias in contingent valuation: Analysis of a combination phone/mail survey. *Economics Letters* 41:215-20.

Whitehead, John C., Peter A. Groothuis, Thomas J. Hoban, and William B. Clifford. 1994. Sample bias in contingent valuation: A comparison of the correction methods. *Leisure Sciences* 16:249-58.

Whittington, Dale. 1998. Administering contingent valuation surveys in developing countries. *World Development* 26:21-30.

Whittington, Dale. 2004. Ethical issues with contingent valuation surveys in developing countries: A note on informed consent and other concerns. *Environmental and Resource Economics* 28:507-15.

Willis, Gordon B. 2005. *Cognitive interviewing. A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Windle, Jill, and John Rolfe. 2011. Comparing responses from internet and paper-based collection methods in more complex stated preference environmental valuation surveys. *Economic Analysis and Policy* 41:83-97.

Yan, Ting, and Roger Tourangeau. 2008. Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology* 22:51-68.

Yao, Richard T., Riccardo Scarpa, John M. Rose, and James A. Turner. 2015. Experimental design criteria and their behavioural efficiency: An evaluation in the field. *Environmental and Resource Economics* 62:433-55.

Yuan, Yuan, Kevin J. Boyle, and Wen You. 2015. Sample selection, individual heterogeneity, in valuing farmland conservation easements. *Land Economics* 91:627-49.

Zhang, Jing, and Wiktor L. Adamowicz. 2011. Unraveling the choice format effect: A context-dependent random utility model. *Land Economics* 87:730-43.