Re-examining empirical evidence on contingent valuation – Importance of incentive compatibility

Ewa Zawojska[1], Mikołaj Czajkowski[2]

Abstract

The contingent valuation (CV) method uses respondents' stated choices made in hypothetical situations to infer their preferences for environmental public goods. It enables the general public's preferences to be stated in monetary terms, and hence, to estimate the economic value of a change in the quantity or quality of the goods. However, a key question remains regarding CV's validity: do the value estimates obtained from a CV study reflect respondents' true preferences and their maximum willingness to pay? Numerous empirical investigations have tested CV's validity, but overall conclusions are mixed. We critically re-evaluate this evidence considering the issue of incentive compatibility in contingent valuation settings for which the necessary conditions were recently proposed by Carson and Groves (2007). Our analysis shows that once incentive compatibility conditions are taken into account, the available studies consistently show that the CV method is valid. As a result, we argue that contingent scenarios and elicitation formats must be made incentive compatible in order to observe consumers' true preferences.

Highlights:

- We critically review empirical evidence regarding the validity of contingent valuation
- We assess validity tests considering the incentive compatibility theory
- Studies that satisfy incentive compatibility conditions consistently pass validity tests
- Incentive compatibility is crucial for accurate preference revelation

Keywords: contingent valuation, stated preference, validity, incentive compatibility

JEL Codes: Q51, H4, D6

[1] University of Warsaw, Department of Economic Sciences, Poland, ezawojska@wne.uw.edu.pl

[2] University of Warsaw, Department of Economic Sciences, Poland, miq@wne.uw.edu.pl

1. Introduction

Stated preference (SP) data is commonly collected in surveys to enable researchers to model consumers' preferences, and thus, to determine valuations for the goods or policies under investigation.[3] This process is called contingent valuation (CV), because respondents make choices *contingent* on the hypothetical scenario presented to them in the survey. SP data are particularly useful for valuations of states which are not currently taking place, and for valuations of goods for which no market exists. Therefore, such hypothetical, nonmarket valuations can be essential for effective management and distribution of many environmental and other public goods (Carson and Czajkowski, 2014).

The SP methods are crucial for the efficacious management of goods and allocation of resources. However, the credibility of data obtained from SP methods remains controversial. The method's reliance on respondents' statements, rather than actual market behavior, casts doubt on whether it, in fact, provides an insight into respondents' true preferences. These concerns are supported by the mixed conclusions reported by numerous studies that have tested the validity of SP methods through a range of approaches, often observing a discrepancy between SP responses and real market decisions (e.g., List and Gallet, 2001; Little and Berrens, 2004; Murphy *et al.*, 2005a).

In the face of serious concerns regarding SP's validity on one hand and a great need for effective consumer preference modeling on the other, Carson and Groves (2007) suggested that the observed discrepancy may result from a lack of incentive compatibility in some SP studies[4] and proposed necessary conditions for truthful preference revelation. These conditions include (1) the use of a binary choice survey format, which discourages strategic misrepresentation, and (2) consequentiality, which means respondents believe that their choices in a survey might have consequences in real life. According to Carson and Groves (2007), both conditions need to be satisfied in order to obtain credible SP data.

---

[3] Applications of survey-based methods to determine economic preferences are common not only in environmental economics (Kanninen, 2007), but also in marketing (Louviere *et al.*, 2006), transport (Hensher *et al.*, 2005), health (Nocera *et al.*, 2003), culture (Choi *et al.*, 2010), and many other fields.

[4] Incentive compatibility means that a respondent's optimal strategy is to answer the CV survey truthfully.

In this study, we conduct a critical review of the empirical studies devoted to testing the validity of SP methods. We examine the utilized validity-testing methodologies considering the incentive compatibility theory. We show that once the incentive compatibility conditions are considered, the available studies consistently confirm the validity of SP methods. Accordingly, we argue that it is crucial to make contingent scenarios incentive compatible in order to observe consumers' true preferences.

The remainder of the paper is structured as follows. The next section provides a theoretical background by describing SP techniques, common elicitation formats, the validity tests proposed in the literature, and the necessary conditions for incentive compatibility. Section 3 presents the available empirical evidence regarding the validity or nullification of SP methods, which we critically assess considering the incentive compatibility theory. We conclude by providing a summary of our findings and indicating the areas for future research that we believe have the most potential to make future SP studies accurately reveal respondents' preferences.

2. Theoretical background – Stated preference methods and their validity

2.1. Stated preference methods

Consumers' economic valuation of public goods is often difficult to assess, because no market for these goods exists and consumers' actual purchase decisions cannot be observed. Thus, their preferences cannot be easily determined. This difficulty has spurred the development of methods to determine valuations of nonmarket goods. Over the years, two groups of techniques for calculating nonmarket valuations have been developed. One infers economic value indirectly via the observation of consumers' actual behavior in related markets, and hence, is said to use revealed preferences (RP). The other approach, using the so-called SPs, is based on respondents' choices made in CV surveys constructed in a particular manner. A typical CV survey asks respondents to state their maximum willingness to pay (WTP), or to choose their most preferred alternative, contingent on a hypothetical scenario presented in the questionnaire.

SP methods are widely used for estimating values of nonmarket goods. Their widespread use results from their two main advantages: flexibility (they allow dealing with goods not yet

available in a market or in reality) and their potential to determine the total economic value of a change in a good's provision, including passive-use value (Carson *et al.*, 2001). As a result, in many cases, SPs are the sole feasible valuation method.

SP surveys apply various preference elicitation formats, which can be classified into one of the two categories: matching methods or discrete choice experiments (DCE) (Carson and Louviere, 2011). In matching methods, respondents indicate a specific number that usually expresses their WTP. Open-ended direct question and a payment card are the most commonly used types of matching methods. The former straightforwardly asks respondents to state their WTP for a certain good, whereas the latter provides respondents with a range of monetary values, from which they select the one representing their maximum WTP. In contrast to matching methods, DCE surveys typically ask respondents to choose their most preferred alternative from a given set. Formats within this category differ with respect to the number of choice tasks and possible response options. Table 1 briefly summarizes commonly used DCE approaches.

Table 1. Typology of DCE formats

		Number of choice alternatives (A)	
		$A = 2$	$A > 2$
Number of choice tasks (CT)	$CT = 1$	Single binary choice	Single multinomial choice
	$CT > 1$	Binary choice sequence	Multinomial choice sequence

We will return to the issue of elicitation formats when discussing incentive compatibility requirements.

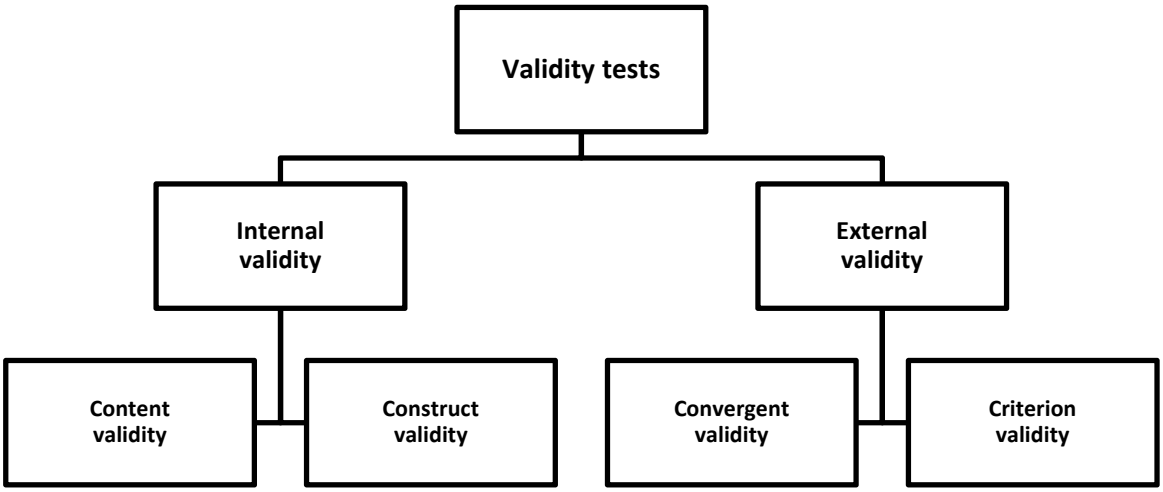
2.2. Validity tests

The issue of SP techniques' ability to provide a valid measure of consumers' true preferences, i.e., whether preferences elicited in surveys accurately reflect true preferences, has been debated for the past few decades. The concept of validity was introduced by Mitchell and Carson (1989), who defined the term as "the degree to which it [the method] measures the theoretical construct under investigation." The authors explained that "the theoretical construct" in the case of CV studies is an individual's WTP, which he/she would definitely pay

in an actual market transaction. Therefore, CV validity tests verify the match between elicited preferences with both theoretical predictions and choices made in real market contexts. True preferences are believed to be revealed in an actual payment setting. Thus, they could be either invoked in an experiment or observed in an actual market, and typically constitute the reference point when comparing elicited SP values.

Many approaches exist to investigate CV validity, which can be grouped into two general categories: *internal* and *external* validity tests. Figure 1 summarizes their typology. Internal validity is most commonly verified using *content* and *construct* validity tests. The former focus on whether the SP survey applies state-of-the-art recommendations of best design practices and often rely solely on the evaluator's subjective opinion. Because of these factors, our study does not discuss content validity tests of SP methods.

Figure 1. Typology of validity tests



The second popular type of internal validity test (construct validity) has been in use since before SP and RP methods became widespread. This type of test assesses CV accuracy by verifying the consistency of observed WTP values with predictions derived from the consumer demand theory, such as sensitivity to price changes, income levels, and other economic variables, which can confirm that responses to CV surveys are not random.

Nevertheless, construct validity tests have two important shortcomings. First, even in real markets, consumers do not necessarily behave in line with the neoclassical demand theory, especially in the case of uncommon goods such as environmental or other public goods, which are usually the subject of SP studies. Given that this theory does not appear to capture all aspects central to consumers' choices, it should be complemented by other concepts such as those provided by behavioral economics. The discrepancy between theoretical predictions and SP values might, therefore, demonstrate not the method's lack of validity, but rather the incompleteness of consumer theory. Second, observing internal consistency does not ensure the coherence of consumers' behavior in an SP study and a real context.

Because of these limitations, external validity tests are usually preferred, and in what follows, we focus on empirical evidence from this category of tests. Within this group, *convergent* and *criterion* validity tests have received the greatest attention. Studies on convergent validity verify the correspondence between WTP estimates derived from an SP survey with some other measure of the same theoretical construct, typically provided by indirect valuation methods. Therefore, this test usually compares value estimates based on a SP study with their counterparts derived from RP approaches such as the hedonic pricing or the travel cost method.

Criterion validity tests investigate behavior consistency in the SP context with choices made in conditions involving actual payments. Unlike convergent validity tests, this approach does not utilize RP-based estimates as a benchmark, but typically elicits consumers' preferences for the same or very similar good in both hypothetical and actual payment settings. The real-payment-based estimates provide a reference point for validity verification.

The consistency between WTP estimates obtained from SP studies and RP or actual payment conditions data is treated as evidence for a method's validity. Conversely, a mismatch between estimated and real values suggests that the method fails to predict consumer behavior. The literature commonly calls such a discrepancy a "hypothetical bias," because it is usually ascribed to the hypothetical nature of SP surveys, which provide different incentives from those experienced in real-life situations. The issue of the incentive properties of SP surveys is, therefore, discussed in detail in the next section.

2.3. Incentive properties of SP surveys and elicitation formats

While validity tests are a useful tool, a more general question arises: when can it be concluded that respondents' preferences revealed in a survey context are the same as the preferences exhibited in real-life situations? The answer is provided by the incentive compatibility theory. If survey choices are incentive compatible, the observed preferences should be consistent with respondents' actual behavior.

Carson and Groves (2007) introduced the necessary conditions for SP surveys to be incentive compatible, and hence, to be able to reveal respondents' true preferences. First, the surveys should be perceived as consequential. To be consequential, a question needs to have two features: participants must care about the problem raised in a survey and they must believe that their responses will influence the agency's final decision. In opposite circumstances, a question is inconsequential.[5]

The second condition determines the elicitation format. A single binary choice question with one alternative being the status quo has long been recognized as the format allowing truthful preference revelation, under the condition that the agency is perceived in force to introduce the proposed alternative (Farquharson, 1969). This format is the subject of much attention in CV studies, which largely results from the recommendations of the National Oceanic and Atmospheric Administration (Arrow *et al.*, 1993). Alternative elicitation questions, i.e., those including more than two alternatives or more than one choice situation, are generically not incentive compatible (Gibbard, 1973; Satterthwaite, 1975), as explained below.

Optimal response strategies for elicitation formats other than those offering a single binary choice typically diverge from truthful preference revelation in favor of strategic misrepresentation. The lack of incentive compatibility arises mainly from respondents' uncertainty regarding how survey votes will be converted into final actions (Carson and Groves, 2007). In the case of a single multinomial choice question, if a respondent is convinced

[5] The crucial role of consequentiality has long been recognized. Hoehn and Randall (1987) emphasize that "a key assumption" underlying the application of SP methods is respondents' conviction about "some influence [of survey results] on the eventual policy decision." Instead, many existing CV studies rely on the so-called "epsilon truthfulness" assumption, according to which a respondent who does not perceive any gain or loss from the way the survey is answered gives truthful responses (Rasmusen, 1989). As innocuous as it appears, this is a very strong assumption, and the need to avoid it has been long been recognized (Kurz, 1974).

that the agency will introduce only one of the proposed options, then in order to influence the final decision, it is rational for a respondent to limit his/her choice possibilities to the two alternatives with the highest probability of winning (much like voting in presidential elections with more than two candidates). It is possible that with more than two alternatives, respondents would exclude their unconditionally most preferred alternative if they find it unlikely to be implemented.

In a sequence of binary choice questions, on the other hand, the desirable incentive properties of a single binary question can only be retained if respondents consider all choice tasks in a sequence independently. Otherwise, they do not answer a particular binary question, but rather place it in the context of choices made in previous choice sets, compare it with the alternatives presented in preceding tasks, and expect that their choices will change future offers that are made (much like in negotiations). Indeed, the repetitive format has been shown to invoke problems such as starting point bias (Herriges and Shogren, 1996) or reference point revision (DeShazo, 2002), which can be considered as resulting from the question format's lack of incentive compatibility.

Finally, the incentive properties of matching methods are doubtful because (1) respondents have no incentive to state their *maximum* WTP and (2) the conditions this method creates are far from market transactions; in reality, consumers do not usually need to define their maximum WTP, but merely decide whether to buy a good at a given price. Indeed, this format often leads to high nonresponse rates and many protest answers, which typically stem from respondents' difficulty in stating a continuous WTP value.[6]

In summary, to elicit accurate preferences, a SP survey must satisfy rather stringent conditions. In addition, as demonstrated in the voting literature (Farquharson, 1969), a CV question needs to feature a take-it-or-leave-it offer, meaning that the respondent's vote is not tied to any other potential offers he/she may get, and moreover, the agency should be able to coercively collect payment for a good if it is provided. Needless to say, many current

[6] The payment card mechanism, which was designed to overcome problems tied to the open-ended direct question, also fails to offer incentive-compatible conditions. This format can be viewed as a form of a single multinomial choice question, which lacks incentive compatibility properties, as discussed before. A stylized fact illustrating the lack of incentive compatibility of a payment card elicitation mechanism is the estimated WTP's dependence on the number and levels of bids used and even their order on a choice card.

empirical applications of SP methods do not satisfy these conditions. Although the extent of deviation caused by violating any of the above conditions still needs to be empirically investigated,[7] it is reasonable to evaluate the SP method's validity only if incentive compatibility conditions are satisfied. In what follows, we critically review existing validity test results available in the literature by placing the empirical evidence in the context of the incentive compatibility theory.

3. Critical review of the available external validity test results

3.1. Convergent validity tests

Convergent validity tests assess how closely WTP estimates derived from SP studies correspond to other measures of economic value obtained in different, typically market-based, methods. Consequently, these tests usually compare value estimates based on SPs with revealed preferences.

The approach was first applied by Knetsch and Davis (1966), who compared SPs toward outdoor recreation in a forest with estimates from a travel cost analysis. Since this study, many researchers have addressed the question of the convergent validity of the CV method, which is well summarized in the meta-analysis by Carson *et al.* (1996). Their investigation of 83 studies, encompassing 616 comparisons of SP estimates to their RP counterparts, shows that in the case of quasi-public goods, stated SP results are somewhat underestimated. The mean ratio of stated to revealed preference estimates is 0.89 with a 95% confidence interval of [0.81; 0.96]. For a weighted sample,[8] the average ratio of 0.92 does not differ significantly from 1.0, which indicates convergent validity.

In contrast to the general investigation of Carson *et al.* (1996), which includes studies devoted to various quasi-public goods, other meta-analyses usually focus on specific categories of goods. Walsh *et al.* (1989; 1992) have conducted the first meta-analyses of recreation

[7] A noteworthy exception is the requirement of consequentiality, as theoretical and empirical evidence of its importance continues to mount (for example, Vossler and Evans, 2009; Broadbent *et al.*, 2010; Herriges *et al.*, 2010; Vossler *et al.*, 2012a; Vossler and Watson, 2013; Carson *et al.*, 2014).

[8] The weighted dataset treats the mean SP to RP ratio from each study as one observation when the study provides multiple estimates.

valuation studies to assess the discrepancy between estimates from CV and travel cost methods. They find that RP estimates usually exceed those derived from SP methods. Similar results are obtained in meta-analyses by Rosenberger and Loomis (2000) and Shrestha and Loomis (2003), who used 682 estimates from 131 studies to perform in-sample and out-of-sample convergent validity tests, respectively. The convergent validity of studies devoted to recreational goods was also investigated by Rolfe and Dyack (2010) and Whitehead *et al.* (2010), who found that SP methods tend to produce lower value estimates than the travel cost method; however, they are statistically equivalent in terms of the predicted number of trips. Ferrini *et al.* (2014) have also reported contradictory evidence, i.e., travel cost and CV payment card estimates do not differ significantly, whereas estimates derived from dichotomous choice appear significantly higher than their counterparts obtained from other methods.

Johnston *et al.* (2006) limit their meta-analysis to the valuation of recreational fishing. Using 391 observations from 48 studies conducted in the years 1977–2001, the authors are unable to draw univocal conclusions regarding the relationship between SP and RP estimates. They find that the effect of the applied methodology (SP or RP) on WTP depends on other characteristics such as the year a study was conducted.

In the context of environmental goods, Foster *et al.* (1997) assess CV validity by comparing actual donations to six large-scale fund collections organized by The Royal Society for the Protection of Birds with the results from several CV studies which address the issue of comparable environmental amenities. Their analysis suggests a significant upward divergence of CV responses in comparison with actual donations.

Woodward and Wui (2000) perform a meta-analysis of 39 studies valuing wetlands. Their findings indicate that SP and RP studies produce inconsistent estimates. CV estimates appear lower than those based on hedonic price methods, but they are not statistically different from the results derived from the travel cost method.[9] Brander *et al.* (2006) extend the research sample used by Woodward and Wui (2000) to 80 studies, which provide 215 observations. Their meta-analysis showed that CV generates significantly higher values than other

[9] These results should be treated with caution because the sample included only two studies that applied the hedonic price method.

techniques, including hedonic pricing and travel cost methods, which essentially contradicts earlier findings. Brander *et al.* (2007) focus on the assessment of coral reefs. Referring to 52 coral reef valuation studies comprising a total of 100 observations, they find that CV methods generate statistically lower value estimates in comparison to other valuation techniques, including the travel cost method.

Some researchers assess CV convergent validity using private goods. Such tests are typically based on the evaluation of how closely consumers' behavior in a real market correspond with results of a CV survey regarding the same good. Shogren *et al.* (1999) compare valuations of chicken breasts, Loomis *et al.* (2000) investigate willingness to pay for elk and deer hunting permits, whereas Alfnes *et al.* (2006) assess the consistency of consumers' real choices and CV responses with regard to a color of salmon. Each of those studies report significant, upward bias of estimates based on a CV survey as compared with actual market behavior. However, the evidence across existing convergent validity tests is not so univocal. For example, Hudson *et al.* (2012) observe that the direction of the bias depends on the good valued – CV-based estimates are significantly higher than in-store actual transactions suggest in the case of prawns and marine shrimps, while an opposite (downward) divergence of CV-based estimates from actual behavior occurs for valuation of lobsters. On the other hand, Lusk *et al.* (2006) report no significant differences between CV responses and in-store transactions.

In transportation, Brownstone and Small (2005) find that travel-time savings observed in DCEs underestimate actual values of travel time spent in congested traffic. Similarly, Fifer *et al.* (2014) report significant differences between preferences for driving distance as stated in a DCE and those revealed during a 10-week GPD driving field study.

In the context of health economics, Kochi *et al.* (2006) conducted a meta-analysis of studies valuing a statistical life. Their examination of 197 observations obtained from 40 studies suggests that CV produces statistically lower estimates than hedonic wage techniques. Clarke (2002) compares SPs toward mammographic screening with the RP derived from a travel cost study. The author observes that the SP method leads to significantly higher WTP estimates than the travel cost technique. He ascribes this finding to potential altruistic attitudes. On the other hand, Kesternich *et al.* (2013) find no significant discrepancy between preferences

expressed in a DCE study on public health insurance programs and actual choices observed in the market.

In summary, convergent validity tests appear to provide mixed conclusions with respect to the accuracy of SP methods. However, we note that one should not consider these results at face value, because the methodology of these comparisons often suffers from serious shortcomings. The crucial limitation follows from differences in application between SP and RP studies. In RP studies, consumer behavior can only be observed with respect to private and quasi-public goods, whereas SP methods are most commonly used in the context of public goods. In addition, RP techniques enable the valuation of goods and services that have actually been provided, whereas SP methods are usually used for the valuation of hypothetical new states and typically consider changes in values rather than actual total values, which further limits the validity of comparisons. The picture is further complicated by the fact that contrary to SP, RP methods cannot capture passive-use values. Finally, SP studies and their RP counterparts often use similar, but not identical, goods or services and target different populations. These differences impose important limitations on the extent to which value estimates derived from RP and SP studies are expected to be equivalent.

In addition to the above reservations, another potential problem that may render SP and RP studies incomparable is the lack of satisfying incentive compatibility conditions, particularly the requirement of the survey's perceived consequentiality. If the SP responses are not collected in incentive-compatible conditions, there is no guarantee that they reflect respondents' true preferences. This provides a yet another reason for the ambiguous results obtained from convergent validity tests. As a result, the comparisons and meta-analyses listed above could be improved by only including empirical studies that adhere to incentive compatibility conditions. Among the reported convergent validity tests, only the study of Alfnes *et al.* (2006) compares actual market behavior with preferences stated in a fully consequential choice experiment. The finding of a significant upward bias of SP with respect to RP is, however, still limited because of the other reservations listed in the preceding paragraph.

Considering all the above limitations, we conclude that the existing comparisons of SP and RP study results might not be the best methods to evaluate the validity of SP methods in a scientifically sound manner. As a result, we now consider criterion validity tests, which have

received more attention in the literature and appear to be more appropriate for verifying the validity of SP methods than convergent validity tests.

3.2. Criterion validity tests

Criterion validity tests potentially offer the most conclusive verification of SP validity (Mitchell and Carson, 1989). They compare respondents' SP choices with different actual payment settings, obtained in actual and simulated market conditions, induced-value lab experiments, or naturally occurring public referenda. In what follows, we review the empirical evidence from criterion validity tests, organized according to these categories.

3.2.1. Actual market studies

Criterion validity tests that use actual market data typically compare values from field CV surveys with consumers' real purchasing behavior. Their unquestionable advantage is the use of field (rather than lab) CV surveys, as they can mostly recreate the usual implementation conditions of SP methods. On the other hand, they are naturally bound to private or club goods, because there are markets in which purchasing behavior can be observed.[10] Although this discrepancy between test procedures being applied to private goods and SP methods usually applied to value public goods does cast some doubt on the applicability of their conclusions, it is nonetheless useful to look at them closer.

An overview of criterion validity studies that use market transactions of private goods is provided in Appendix Table A1. The evidence arising from these studies is mixed; while some studies report the equivalence of results, others indicate that values observed in CV surveys are smaller or larger than those observed in real markets.

Bishop and Heberlein (1979) and Bishop *et al.* (1983) conducted some of the first criterion validity tests of the CV method. Their field investigations suggest downward hypothetical bias, meaning that respondents in CV surveys usually report lower values than implied by their

[10] Another point raised in the support of criterion validity tests based on actual market studies is respondents' familiarity with the good. This, however, is not required by any of the incentive compatibility conditions.

actual market behavior. Although at the first glance these results contest CV validity, their outcomes appear consistent with predictions based on rational choice theory, i.e., when an existing private good is considered, rational agents may intentionally report lower values if they think that the survey will be used for pricing purposes. Stating a lower WTP could, thus, be seen as a tool to avoid future increase in prices.

On the other hand, some studies (for example, List and Shogren, 1998; Loomis *et al.*, 2009; Chowdhury *et al.*, 2011) find a positive difference, indicating that people state higher values than those they are actually willing to pay. However, this evidence can also be discredited because such behavior is justified on the basis of a survey's incentive structure. If a respondent believes that there is a nonzero probability that he might want to buy a product at the stated price (now or in future), he should declare high WTP to increase the probability of making the good available in the market. After all, once the good is delivered, the purchasing decision can always be reconsidered.

Obtaining different WTP estimates in SP surveys compared to actual market studies is not a rule. For example, Dickie *et al.* (1987) and Smith and Mansfield (1998) do not observe statistically significant discrepancies between SP and market-based values. Considering the above reservations, it is possible that the two effects cancel out (on average), and hence, such results are not sufficient to determine CV's validity. It is, therefore, necessary to consider the incentives that respondents experience.

Moreover, studies reported above are not consequential, which alone offers sufficient grounds to question their conclusions. The evidence following from some of those studies suggests that respondents' perceptions on consequentiality might impinge on the observed discrepancy between CV and actual choices. Limiting the sample only to respondents definitely certain about their choice, as done by Blumenschein *et al.* (2001), or using CV surveys with cheap talk,[11] as done by List *et al.* (2006), lead to consistency of respondents' stated and actual choices. The effect of cheap talk scripts is ambiguous, however – List (2001) reported that the significant hypothetical bias was eliminated due to the use of cheap talk

[11] Cheap talk provides respondents with additional information before the actual valuation question. It reminds the agents about the hypothetical survey character and directly discusses the impact of hypothetical bias on self-reported values.

only for unexperienced market participant, while Moser *et al.* (2014) observed that the bias still exists even when cheap talk scripts are included.

Carson *et al.* (2014) and Landry and List (2007), summarized in the Appendix Table A2, compare SP and actual market choices in a consequential setting.[12] Although the authors use private goods, those goods simulate the provision of a public good, because respondents vote in a referendum whether the private goods (sports memoriabilia) should be provided to all voting participants. Referenda used in both studies vary with respect to the levels of the probability of being binding. Carson *et al.* (2014) find that referendum participants who are informed about the positive likelihood of real consequences of voting do not display significantly different behavior across various probability levels (20%, 50%, or 80%). Similarly, Landry and List (2007) observe that cheap talk and consequential treatments with a 50% chance of the referendum being binding lead to consistency between SP WTP estimates and actual choices. At the same time, purely hypothetical voting yields significantly different results than those observed in real contexts. These findings are in line with the predictions of the incentive compatibility theory and illustrate the necessity of consequentiality as a prerequisite of truthful preference revelation.

Donation is another payment mechanism occasionally used in CV studies. In surveys employing this approach, respondents are typically asked about their willingness to contribute to the provision of a public good (for example, Swallow and Woudyalew, 1994; Brown and Duffield, 1995; Loomis and Gonzalez-Caban, 1997). As argued by Champ *et al.* (1997), respondents may see donations as more plausible in certain survey contexts than a tax increase. Several studies test CV criterion validity applying a donation mechanism. The results of these studies are summarized in the Appendix Table A3.

The empirical evidence suggests that respondents' actual voluntary contributions are often overestimated by SP questions (for example, Seip and Strand, 1992; Brown *et al.*, 1996b;

[12] Herriges *et al.* (2010) is a notable example of an out-of-laboratory study conducted in a public good context, contributing to the discussion on the CV's validity, although not being a standard validity test. Instead of comparing consumers' stated choices with behavior in an actual payment setting, Herriges *et al.* (2010) compare preferences of respondents perceiving a CV survey as consequential and non-consequential The authors find that the WTP of respondents who believe (even to the least extent) that the survey could be consequential is significantly different from respondents who perceive the survey as inconsequential. This contributes to the discussion on the role of consequentiality for truthful preference revelation.]

Brown and Taylor, 2000). This result is in fact expected due to the incentive structure provided by this mechanism. When discussing real donations, rational participants free ride and let the public good be provided through contributions from others. However, in a CV survey it makes sense to overstate a WTP in order to have the good be seen as worth providing (Bateman *et al.*, 2002). The effect is intensified particularly when a good is perceived as socially desirable (cf. purchasing moral satisfaction, Kahneman and Knetsch, 1992). Loomis *et al.* (2009) and Norwood and Lusk (2011) find that the hypothetical context and social desirability intensify the bias.

Although none of the studies investigating criterion validity with the use of donations applies a properly consequential setting, some researchers control how certain respondents are that they will actually behave in a way they stated in a CV survey. Excluding uncertain respondents allows typically to obtain statistically indifferent value estimates from CV surveys and actual choices, as the research by Champ and Bishop (2001), Champ *et al.* (1997), Champ *et al.* (2009) and Ethier *et al.* (2000) show. MacMillan *et al.* (1999) also explain the lack of a significant difference between stated and actual behavior on the grounds of the realistic nature of their study. The authors enhance the survey realism by its association with an actual appeal fund which considerable media attention and obtained wide national prominence. Moreover, similarly as in the case of market studies based on private goods, cheap talk scripts considerably attenuate the discrepancy between stated and actual choices (List *et al.*, 2006; Champ *et al.*, 2009).

In sum, the majority of actual market studies aimed at testing SP validity provide no robust basis to assess the accuracy of CV-based estimates in measuring WTP. Without assuring that incentive compatibility conditions are satisfied, establishing what these studies really test is difficult because the observed discrepancy between SP choices and actual behavior is exactly what is predicted according to the economic theory. On the other hand, very few studies (Landry and List, 2007; Carson *et al.*, 2014) control consequentiality, which is a crucial element of SP studies' incentive compatibility. Remarkably, the studies that do assure incentive compatibility conditions report a close correspondence between SPs and observed behavior.

3.2.2. Simulated market studies

Another possibility to test CV validity is provided by artificial markets created in a laboratory environment. Simulated market studies allow comparisons between respondents' decisions taken in hypothetical settings with their equivalents involving real money payments. The results obtained in actual payment treatments are considered a close measure of individuals' true WTP, and hence, a suitable reference for comparisons. The laboratory setting makes it possible to test the SP methods' validity using different types of goods (private, public and quasi-public), employing various mechanisms to determine the final outcome (donations and referenda), and applying either home-grown, or induced values.[13]

Numerous criterion validity tests in a laboratory setting have been conducted using private goods (Appendix Table A4 provides an overview). Again, the evidence resulting from these studies is mixed and does not allow univocal conclusions to be drawn, however, the great preponderance of studies report a statistically significant upward discrepancy between estimates from hypothetical and actual payment conditions. Nevertheless, we argue that since the essential prerequisite for truthful preference revelation is adherence to incentive compatibility conditions, the available studies do not provide a valuable input to the discussion, because none of them compare estimates from real payment conditions to those obtained in a consequential setting, i.e., the hypothetical treatments used by these studies do not suggest any consequences arising from the survey questions. In fact, these experiments have designs aimed at creating purely hypothetical conditions.[14] As a result, considering the necessary conditions for incentive compatibility, these experiments' results do not shed much light on the issue of the validity of state-of-the-art SP methods.

Because private goods are not typically the subject of CV studies, the results of experiments dealing with public goods are more informative. An overview of such studies is available in the Appendix Table A5. Most show a significant divergence between respondents' behavior in

[13] In induced-value experiments, a researcher ascribes values to specific experimental actions or results, which are presented to subjects in the instructions (e.g., a payoff to all participants if they jointly satisfy a condition). These actions or results have no value in and of themselves. In contrast, home-grown value experiments elicit agents' personal preferences, which they bring with them to the experiment, and thus, existed prior to the experiment, not ones introduced by the experimental setting.

[14] A possible exception is the experiment reported by Johannesson *et al.* (1998), who emphasize that respondents should provide the amount they would pay "here and now" in contrast to the usual "would you ever pay..."

hypothetical and actual settings. Again, however, these findings do not incorporate the necessary conditions for incentive compatibility, and hence, respondents' answers cannot be used without reservation as straightforward reflections of their true preferences.

Among simulated market studies on public goods applying the donation mechanism, only Broadbent *et al.* (2010) and Broadbent (2012) consider the role of consequentiality. They find that respondents believing in survey consequentiality[15] declare statistically different values in hypothetical and actual payment settings. However, caution should be exercised when interpreting this result. First, a surprisingly large share of research participants viewed the survey as binding, which could be tied to the binary (yes-or-no) format of the follow-up question aimed at verifying perceived consequentiality.[16] More importantly, these studies apply a donation vehicle that does not create incentive compatible conditions, because it is not coercive. Finally, we note that the experimental sample was possibly mismatched with the contingent scenario, e.g., early-year students were asked to contribute to a program of trail development (Broadbent, 2012).

To the best of our knowledge, the only studies using a referendum format[17] and complying with the consequentiality requirement are reported by Cummings and Taylor (1998), Vossler and Evans (2009), and Vossler *et al.* (2012b). Cummings and Taylor (1998) implement treatments that vary the odds of a referendum being binding. The probability range encompasses referenda with 0%, 25%, 50%, 75%, and 100% chances that the vote cast in the survey will lead to real consequences. The authors find that a relatively high probability (exceeding 50%) is required for respondents to state behavior that is statistically indistinguishable from their actual choices. Although the referenda that assigned 25% and 50% chances of being binding had lowered shares of "yes" WTP responses, when compared

[15] Personal perceptions of consequentiality are typically measured through self-reports to a follow-up question regarding how strongly a respondent believes in real consequences resulting from a CV survey.

[16] In contrast, Herriges *et al.* (2010) measure consequentiality perception through self-reports on a five-degree Likert scale, Hwang *et al.* (2014) use a four-degree scale, and Vossler *et al.* (2012b) assess consequentiality perceptions on a six-degree scale.

[17] A referendum format is more likely to be incentive compatible, as it clearly states the provision rule (usually voting), and is typically linked with a payment mechanism such as a tax increase, which excludes the possibility of free riding.

to a purely hypothetical treatment, the results remained different from the real (100% binding) referendum.

Vossler and Evans (2009) examine respondents' behavior in hypothetical, advisory, and binding referenda. They find that responses provided in advisory referenda are sincere as long as an unknown or explicit but modest weight is put on their votes. A relatively small influence on participants' responses, however, results in WTP estimates similar to those from a purely hypothetical survey. This suggests that the lack of precise information about how survey results will be used by policy makers does not necessarily bias the results, provided that respondents are assured about the survey's consequentiality (influence).

Vossler *et al.* (2012b) arrive to similar conclusions while relying on respondents' self-perceived consequentiality (in contrast to Vossler and Evans (2009), who objectively defined consequentiality through survey scripts). When the stated and real WTP functions are estimated only for participants who believe that the survey results would have any influence on policy, no statistically significant discrepancy is found. Overall, this result corroborates the evidence found in other incentive-compatible studies, which consistently indicate the validity of the SP methods.

Finally, instead of eliciting respondents' personal preferences (*home-grown* values with which they come to the experiment), a different stream of research aimed at verifying the validity of SP methods using induced-value experiments is performed, in which a researcher defines payoffs associated with particular outcomes and subsequently analyzes whether respondents behave in line with the induced preferences. This approach is considered a clear test of the consistency of survey responses with respondents' true values; any deviation can be easily detected because true preferences are known to the experimenter. This answers a broadly raised objection that researchers do not know survey participants' actual preferences, and hence, the study's coherence cannot be verified. Furthermore, induced-value experiments exclude any potential bias related to the type of a good used in a study because preferences are determined in the abstract context, solely on the basis of monetary values.

Appendix Table A6 presents an overview of the results. The preponderance of induced-value experiments suggests the validity of SP methods by revealing consistency between reported

and actual preferences. In fact, if only consequential studies are considered,[18] the conclusion becomes even stronger, i.e., no significant discrepancy between stated and true values exists.

3.2.3. Naturally occurring referenda studies

Despite the advantages of verifying CV validity through simulated market studies, which include the possibility to control the research conditions to the greatest extent, the literature raises doubts regarding the reliability of this type of test (Taylor, 1998; Bateman *et al.*, 2002; Poe *et al.*, 2002). Opponents argue that the simulated market environment does not appropriately capture true incentives operating in a real context, and thus, consumers' behavior cannot be translated into the actual application of SP methods. In response to this objection, some researchers utilize naturally occurring referenda.

It is often claimed that actual voting behavior might provide the most accurate reference point for SP validity testing. Arrow *et al.* (1993) stress that "a critically important contribution could come from experiments in which state-of-the-art CV studies are employed in contexts where they can in fact be compared with 'real' behavioral willingness to pay for goods that can actually be bought and sold." Naturally occurring referenda offer such a possibility, because they elicit preferences in a context free from the experimental setting.

Appendix Table A7 summarizes the results of studies based on naturally occurring referenda. With one exception, all such available studies support the SP method's validity. Some of these, however, may depend on how the "undecided" votes to a binary choice question are treated. Carson *et al.* (1987) treat 60% of the "undecided" as "no" votes, whereas Champ and Brown (1997) and Vossler *et al.* (2003) do so for all undecided votes. On the other hand, the survey participants might have expected or known about the upcoming public referendum and already made up their minds regarding their votes, thus inflating the similarity between a

[18] Carson *et al.* (2009) investigate double referenda and report that even if this mechanism does not fully meet incentive-compatibility requirements (that is, when the two binary questions in the sequence are not perceived as being independent), value estimates, although biased, do not diverge much from true preferences. Collins and Vossler (2009) observe a very low frequency of deviations from induced values. Mitani and Flores (2012) find strong support for coherence between voting and underlying preferences. At the same time, considering various probabilities of a referendum being binding (1%, 10%, and 25%), the authors observe that the lower the probability of consequentiality, the higher is the frequency of deviations. Polomé (2003) shows that individually reported values strongly correlate with induced preferences.

survey and actual votes. Schläpfer *et al.* (2004) argue that only SP surveys that are conducted before an actual referendum is discussed or announced to the public should be used to gauge their validity.

The only two studies that satisfy Schläpfer's stipulation and could have been perceived as consequential are Johnston (2006) and Vossler and Watson (2013). Johnston (2006) compares respondents' behavior in a CV survey on the provision of a quasi-public good with that in a subsequent real referendum and finds no significant difference. The survey, which preceded actual voting, was consequential because it determined whether the real referendum would occur. Vossler and Watson (2013) obtain similar results. They find that SP studies under-predict the number of referendum votes in favor of the program, but once only the respondents who perceive the survey as consequential are included; no divergence between hypothetical and real choices occurs.

Overall, the evidence from naturally occurring referenda also adds support to the validity of SP methods, as long as the incentive compatibility conditions are considered.

4. Summary and conclusions

The issue of the SP methods' validity has been broadly investigated, particularly because the empirical evidence is often contradictory: some studies report significant differences between stated and true preferences, whereas others provide support for the CV methods' validity. Our review sheds new light on the issue by critically evaluating the existing empirical evidence considering the incentive compatibility theory. We argue that the mixed evidence can be explained by determining whether a study adheres to the necessary conditions of incentive compatibility (Carson and Groves, 2007). When the available studies are limited only to those that satisfy these requirements, the evidence becomes univocal – respondents' stated and true preferences are the same.

We critically reviewed four main approaches to test the validity – content, construct, convergent, and criterion validity – highlighting their strengths and weaknesses. We argue that criterion validity is the most adequate and thus placed most emphasis on the results of these studies. By classifying the empirical evidence with respect to whether it (1) deals with private or public goods, (2) uses a coercive or voluntary payment mechanism, (3) can be

perceived by respondents as consequential, and (4) uses a single binary choice format, we could identify studies that provide meaningful results in terms of providing conditions in which rational respondents are actually expected by economic theory to answer in line with their true preferences. The results of such studies consistently point to the validity of preferences stated under such conditions.

Our review indicates a few promising directions for future analyses. First, although the overwhelming majority of CV validity studies are performed in labs, very little is known regarding the direct applicability of such evidence to real-life situations.[19] Field experiments could shed more light both on their validity in general and on whether lab tests of CV indeed provide valuable input, providing that those field experiments adhere to incentive compatibility conditions. Second, research on consequentiality perception is still in its infancy (Kling *et al.*, 2012). Crucial questions concern measuring the perceived level of consequentiality (Nepal *et al.*, 2009; Herriges *et al.*, 2010) and making the measurements themselves incentive compatible. Finally, a relatively small body of empirical literature addresses the problem of the bias resulting from using more than two alternatives or more than one choice task per respondent. If the resulting bias is relatively small, the statistical efficiency of *some* more elaborate elicitation formats could outweigh the bias resulting from being theoretically incentive incompatible. We believe these issues provide an opportunity for one of the most valuable contributions to the field of SP methods.

Over 50 years of empirical experience concerning the implementation of various SP methods has accrued, and much has been learned. One of the main areas where SP methods have matured lies in understanding the effects of the conditions in which respondents make choices. This learning process resulted from the necessity to address criticism and explain various anomalies observed in some variants of the method and eventually confirmed the need for incentive compatibility. We show that once the conditions for incentive compatibility are considered, SP methods appear to consistently provide valid estimates of consumer preferences. This result is reassuring and indicates that if designed and conducted

[19] For example, lab data is usually collected from respondents who are aware of the fact that they are participating in the experiment.

appropriately, SP methods will remain of central importance to modern welfare economics and environmental economics in particular.

Appendix A1. Criterion validity tests in actual private good market studies

Author	Good	Elicitation mode	Elicitation format[20]	Sample	Divergence[21]
Bishop and Heberlein (1979)	goose hunting permits	mail	SBC followed by OE	split	significant, downward
Bishop *et al.* (1983)	goose hunting permits	mail	SBC followed by OE	split	significant, downward
Blumenschein *et al.* (2001)	asthma management program	in-person interview at a pharmacy preceded by a phone invitation	SBC	split	significant, upward; insignificant if limited to "definitely yes" responses
Chang *et al.* (2009)	dishwashing liquid, ground beef, wheat flour	field group session preceded by an invitation	M-SEQ	split	significant, upward
Chowdhury *et al.* (2011)	biofortified sweet potatos	field survey preceded by an invitation	M-SEQ	split	significant, upward
Dickie *et al.* (1987)	fresh strawberries	in-person interview	OE	split	insignificant
List (2001)	baseball cards	field intercept survey (at a sports-card show)	second-price auction	split	significant upward; insignificant for unexperienced market participants answering surveys with cheap talk
List *et al.* (2006)	sports cards	field intercept survey (at a sports-card show)	M-SEQ	split	significant, upward; insignificant for a survey with cheap talk
List and Shogren (1998)	baseball cards	field intercept survey (at a sports-card show)	second-price auction	within	significant, upward
Loomis *et al.* (2009)	bottled water protecting infant health	in-person interview, mail	M-SEQ	split	significant, upward
Moser *et al.* (2014)	apples	CAPI (in supermarkets)	M-SEQ	split	significant, upward for surveys with and without cheap talk
Shogren *et al.* (1999)	chicken breasts	mail (CV), field group session preceded by an invitation (actual), phone-mail-phone interview	DBQ (CV, actual), PC (actual)	split	significant, upward
Smith and Mansfield (1998)	opportunity cost of survey participation		SBC	split	insignificant
Yue and Tong (2009)	organic and locally grown tomatoes	field survey	M-SEQ	split	significant, upward

[20] The following notation is used (applies to all tables in the Appendix):
SBQ – a single binary choice question, DBQ – a double bounded binary question, MBQ – a multiple-bounded question, B-SEQ – a binary choice sequence, SMC – a single multinomial choice question, M-SEQ – a multinomial choice sequence, OE – an open-ended direct question, PC – a payment card, PL – a payment ladder (a variation of PC with ordered values).

[21] "Divergence" expresses the divergence of CV-based estimates from actual values. Upward divergence means that CV-based value estimates exceed those derived from an actual payment setting.

Appendix A2. Criterion validity tests in actual public good market studies using referenda

Author	Good	Elicitation mode	Elicitation format	Sample	Divergence
Carson *et al.* (2014)	baseball memorabilia	field referenda with 0%, 20%, 50%, 80% and 100% probability of being binding	SBC	split	significant, upward for a 0% probabilistic referendum; insignificant for 20%, 50% and 80% probabilistic referenda
Landry and List (2007)	sports memorabilia	field referenda with cheap talk or with 0%, 50% and 100% probability of being binding	B-SEQ (double)	split	significant, upward for a 0% probabilistic referendum; insignificant for a referendum with cheap talk and for a 50% probabilistic referendum

Appendix A3. Criterion validity tests in actual public good market studies using donations

Author	Good	Elicitation mode	Elicitation format	Sample	Divergence
Brown *et al.* (1996a)	abandoned road removal from the Grand Canyon National Park	mail	SBC, OE	split	significant, upward
Brown and Taylor (2000)	the Nature Conservancy's rainforest project	in-person interview	OE	split	significant, upward
Cameron *et al.* (2002)	tree planting, providing energy from renewable sources	mail (CV), phone (actual)	SBC (CV, actual), only actual: OE, PC, MBQ, SMC	split	insignificant for SBC, SMC; significant, upward for OE and MBQ; significant for PC (upward for median WTP, downward for mean WTP)
Champ and Bishop (2001)	wind-generated electricity	mail	SBC	split	significant, upward; insignificant if "uncertain" responses excluded
Champ *et al.* (1997)	abandoned road removal from the Grand Canyon National Park	mail	SBC	split	significant, upward; insignificant if "uncertain" responses excluded
Champ *et al.* (2009)	purchase of radio transmitters for whooping cranes	mail	SBC	split	significant upward (mitigated by cheap talk); insignificant if limited to "certain" responses
Duffield and Patterson (1992)	contribution to Montana Leasing Trust Fund	mail	PC	split	insignificant if the CV survey is sent under the university letterhead; significant upward if the CV survey is sent under the Nature Conservancy letterhead
Ethier *et al.* (2000)	green-pricing project	phone (CV, actual), mail (CV)	SBC	split	significant, upward; insignificant if "uncertain" responses excluded
List *et al.* (2006)	a new Centre for Environmental Policy Analysis	mail	SMC	split	significant, upward; insignificant for a survey with cheap talk
MacMillan *et al.* (1999)	purchase and development of the Isle of Eigg	mail	OE	split	insignificant
Poe *et al.* (2002)	renewable energy, tree planting	phone	SBC (CV, actual), OE (CV)	split	significant, upward for SBC; insignificant for OE
Seip and Strand (1992)	membership in an environmentalist association	in-home survey (CV), mail (actual),	SBC	within	significant, upward
Veisten and Navrud (2006)	leasing of virgin forests	mail	SBC, OE	within	significant upward (simultaneous provision of an actual choice survey mitigates the bias)

Appendix A4. Criterion validity tests in simulated private good market studies

Author	Good	Elicitation format	Sample	Divergence
Alfnes *et al.* (2010)	apples	fourth-price auction over 12 alternatives	split, within	significant, upward for surveys with and without cheap talk (information about follow-up actual bidding reduces the bias)
Aoki *et al.* (2010)	avoiding sodium nitrite in ham sandwiches	B-SEQ (forced choice)	split	significant, upward
Balistreri *et al.* (2001)	insurance policy	OE (CV), SBC (CV), English auction (actual)	split	significant, upward for SBC; insignificant for OE
Camacho-Cuena *et al.* (2004)	eco-table with improved recyclability properties	a variant of PL	within	insignificant
Carlson (2000)	t-shirt	second-price auction (actual), OE (CV), SMC (alternatives indicate choice certainty)	within	significant, upward for OE and for "definitely certain" and "probably certain" responses in SMC; insignificant for "definitely certain" respones in SMC
Cummings *et al.* (1995)	electric juicer, chocolate truffles, solar calculator	SBC	split, within	significant, upward
Frykblom (1997)	environmental atlas	SBC, OE	split	significant, upward
Isacsson (2007)	bus ticket	SBC	split	significant, downward
Johannesson (1997)	box of Belgian chocolates	OE (CV), second-price auction (actual)	split	significant, upward
Johannesson *et al.* (1998)	box of Belgian chocolates	SBC	split, within	significant, upward; significant downward if limited to "definitely yes" responses
Kealy *et al.* (1988)	chocolate bar	SBC followed by OE	split	significant, upward
Loomis *et al.* (1997)	art print	SBC, OE	split	significant, upward
Lusk and Schroeder (2004)	beef ribeye steaks	M-SEQ	split	significant, upward
Murphy *et al.* (2010)	coffee mug	PL	split	significant, upward
Neill *et al.* (1994)	watercolor painting, 16th century map of the world	OE (CV), second-price auction (CV, actual)	split	significant, upward
Paradiso and Trisorio (2001)	antique print	OE (CV), second-price auction (actual)	split	significant, upward
Stefani and Scarpa (2009)	weather forecast	SBC	split	significant, upward
Taylor *et al.* (2010)	t-shirt	M-SEQ	split	insignificant
Volinskiy *et al.* (2011)	avoiding the use of genetically modified plants in canola oil	B-SEQ, M-SEQ	split	significant, upward

Appendix A5. Criterion validity tests in simulated public good market studies

Author	Good	Elicitation format	Choice setting	Sample	Divergence
Botelho and Pinto (2002)	environmental educational program	OE	Donation	split	significant, upward
Broadbent (2012)	trail extension plan	M-SEQ	Donation	split	significant (lack of preference equality based on a LR test); insignificant for marginal WTP
Broadbent (2012)	trail extension plan	M-SEQ	Donation	split	significant (lack of preference equality based on a LR test); significant, downward for surveys with cheap talk and if calibrated for certainty; insignificant for marginal WTP
Broadbent *et al.* (2010)	riparian forest restoration	M-SEQ	Donation	split	significant, upward
Carlsson *et al.* (2010)	charity support	B-SEQ (no opt-out option)	Donation	split	significant, upward for surveys with and without cheap talk
Carlsson and Martinsson (2001)	environmental project	B-SEQ (no opt-out option)	Donation	within	insignificant
Getzner (2000)	ibex protection	OE (CV),SBC (CV, actual)	Donation	within	significant, upward
Johansson-Stenman and Svedsäter (2008)	environmental project	B-SEQ (no opt-out option)	Donation	split, within	significant, upward
Murphy *et al.* (2005b)	placing signs to mark trails and rare and endangered species	SBC	Donation	split	significant, upward for surveys with and without cheap talk
Ready *et al.* (2010)	wildlife rehabilitation	M-SEQ	Donation	split, within	significant, upward; insignificant if calibrated for certainty
Sinden (1988)	soil conservation	B-SEQ followed by OE	Donation	within	insignificant
Spencer *et al.* (1998)	water quality monitoring in ponds	SMC	Donation	split	insignificant
Bjornstad *et al.* (1997)	distributing a citizen's guide about groundwater contamination; natural	SBC B-SEQ	Referendum	split, within	significant, upward for "split"; insignificant for "within"

Cummings *et al.* (1997)	environment protection distributing a citizen's guide about groundwater contamination	SBC	Referendum	split	significant, upward
Cummings and Taylor (1998)	distributing a citizen's guide about groundwater contamination	SBC	Referenda with 0%, 25%, 50%, 75% and 100% probability of being binding	split	significant, upward
Cummings and Taylor (1999)	distributing a citizen's guide about groundwater contamination; natural environment protection; rainforest protection; finishing a greenway	SBC	Referendum	split	significant, upward; insignificant for surveys with cheap talk
Krawczyk (2012)	forest protection and restoration	PL	Referendum	split, within	significant, upward
Murphy *et al.* (2010)	a flock of chickens for needy families	PL	Referendum	split	significant, upward
Stefani and Scarpa (2009)	weather forecast	SBC	Referendum	split	insignificant
Taylor (1998)	distributing a citizen's guide about groundwater contamination	SBC	Referendum	split	significant, upward
Vossler *et al.* (2012b)	tree planting	B-SEQ	Referendum	split	insignificant
Vossler and Evans (2009)	on-campus, classroom recycling container	SBC	Referendum	split	significant, upward; insignificant if unknown, or explicit and modest weights are put on respondents votes

Appendix A6. Criterion validity tests using induced-value experiments

Author	Elicitation format	Choice setting	Sample	Divergence
Burton *et al.* (2007)	SBC	Referendum	split, within	significant, upward
Carson *et al.* (2009)	B-SEQ	Referendum	split	insignificant
Collins and Vossler (2009)	B-SEQ, M-SEQ	(Advisory) referendum	split	insignificant
Mitani and Flores (2009)	OE	Donation	within	insignificant
Mitani and Flores (2012)	B-SEQ	Referendum	within	insignificant for consequential referenda; significant violations for low consequential referenda and a small value-cost spread
Mozumder and Berrens (2007)	B-SEQ	Referendum	split	significant, upward; insignificant for surveys with cheap talk
Murphy *et al.* (2010)	a sequence of PL	Becker-DeGroot-Marschak mechanism	split	insignificant
Polomé (2003)	a sequence of PL	(Advisory) referendum; Mean-rule voting (good provided if a mean declaration is at least equal to the cost)	---	insignificant (declarations correlated with induced values); significant for the mean-rule voting (a high induced value increases the probability to overstate; in line with the mechanism incentive properties)
Taylor *et al.* (2001)	SBC	Referendum	split	insignificant
Vossler and McKee (2006)	SBC, PC, PL, MBQ	Referendum (SBC), Random Price Voting Mechanism (PC and MBQ)	split	insignificant

Appendix A7. Criterion validity tests based on naturally occurring referenda

Author	Mode of a CV survey	Elicitation format	Divergence
Carson *et al.* (1987)	phone	SBC	insignificant if 60% of undecided responses treated as "no"
Champ and Brown (1997)	n/a	n/a	insignificant if undecided responses treated as "no"
Johnston (2006)	mail	SBC	insignificant
Schläpfer *et al.* (2004)	phone	B-SEQ	significant, upward
Vossler and Kerkvliet (2003)	mail	SBC, SMC (alternatives indicate choice certainty)	insignificant
Vossler *et al.* (2003)	phone	SBC	insignificant if undecided responses treated as "no"; significant, upward if undecided responses excluded
Vossler and Watson (2013)	mail	SBC	significant, downward; insignificant if limited to respondents perceiving a survey as consequential

References

Alfnes, F., Guttormsen, A., Steine, G., and Kolstad, K., 2006. Consumers' Willingness to Pay for the Color of Salmon: A Choice Experiment with Real Economic Incentives. *American Journal of Agricultural Economics*, 88(4):1050-1061.

Alfnes, F., Yue, C., and Jensen, H. H., 2010. Cognitive Dissonance As a Means of Reducing Hypothetical Bias. *European Review of Agricultural Economics*, 37(2):147-163.

Aoki, K., Shen, J., and Tatsuyoshi, S., 2010. Consumer Reaction to Information on Food Additives: Evidence From an Eating Experiment and a Field Survey. *Journal of Economic Behavior and Organization*, 73(3):433-438.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., and Schuman, H., 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58:4601-4614.

Balistreri, E., McClelland, G., Poe, G. L., and Schulze, W., 2001. Can Hypothetical Questions Reveal True Values? A Laboratory Comparison of Dichotomous Choice and Open-Ended Contingent Values with Auction Values. *Environmental and Resource Economics*, 18(3):275-292.

Bateman, I. J., Carson, R. T., Day, B., Haneman, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D. W., Sugden, R., and Swanson, J., 2002. Economic Valuation with Stated Preference Techniques: A Manual. Edward Elgar, Northampton, Massachusetts.

Bishop, R., and Heberlein, T. A., 1979. Measuring Values of Extramarket Goods: Are Indirect Measures Biased? *American Journal of Agricultural Economics*, 61(5):926-930.

Bishop, R., Heberlein, T. A., and Kealy, M. J., 1983. Contingent Valuation of Environmental Assets: Comparisons with a Simulated Market. *Natural Resources Journal*, 23(3):619-634.

Bjornstad, D., Cummings, R. G., and Osborne, L., 1997. A Learning Design for Reducing Hypothetical Bias in the Contingent Valuation Method. *Environmental and Resource Economics*, 10(3):207-221.

Blumenschein, K., Johannesson, M., Yokoyama, K. K., and Freeman, P., 2001. Hypothetical versus Real Willingness to Pay in the Health Care Sector: Results from a Field Experiment. *Journal of Health Economics*, 20(3):441-457.

Botelho, A., and Pinto, L. C., 2002. Hypothetical, Real, and Predicted Real Willingness to Pay in Open-Ended Surveys: Experimental Results. *Applied Economic Letters*, 9(15):993-996.

Brander, L. M., Florax, R. J. G. M., and Vermaat, J. E., 2006. The Cmpirics of Wetland Valuation: A Comprehensive Summary and Meta-Analysis of the Literature. *Environmental and Resource Economics*, 33(2):223–250.

Brander, L. M., Van Beukering, P., and Cesar, H. S., 2007. The Recreational Value of Coral Reefs: A Meta-Analysis. *Ecological Economics*, 63(1):209-218.

Broadbent, C. D., 2012. Hypothetical Bias, Consequentiality and Choice Experiments. *Economics Bulletin*, 32(3):2490-2499.

Broadbent, C. D., Grandy, J. B., and Berrens, R., 2010. Testing for Hypothetical Bias in a Choice Experiment Using a Local Public Good: Riparian Forest Restoration *International Journal of Ecological Economics and Statistics*, 19(F10):1-19.

Brown, T., Champ, P., Bishop, R., and McCollum, D., 1996a. Which Response Format Reveals the Truth About Donations to a Public Good? *Land Economics*, 72:152-166.

Brown, T., and Duffield, J. W., 1995. Testing Part-Whole Valuation Effects in Contingent Valuation of Instream Flow Protection. *Water Resources Research*, 31(9):2341-2351.

Brown, T., and Taylor, L. O., 2000. Do As You Say, Say As You Do: Evidence on Gender Differences in Actual and Stated Contributions to Public Goods. *Journal of Economic Behavior and Organization*, 43(1):127-139.

Brown, T. C., Champ, P. A., Bishop, R. C., and McCollum, D. W., 1996b. Which Response Format Reveals the Truth about Donations to a Public Good? *Land Economics*, 72(2):152-166.

Brownstone, D., and Small, K. A., 2005. Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations. *Transportation Research Part A: Policy and Practice*, 39(4):279-293.

Burton, A., Carson, K. S., Chilton, S. M., and Hutchinson, G. W., 2007. Resolving Questions about Bias in Real and Hypothetical Referenda. *Environmental and Resource Economics*, 38(4):513-525.

Camacho-Cuena, E., García-Gallego, A., Georgantzís, N., and Sabater-Grande, G., 2004. An Experimental Validation of Hypothetical WTP for a Recyclable Product. *Environmental and Resource Economics*, 27(3):313-335.

Cameron, T. A., Poe, G. L., Ethier, R. G., and Schulze, W., 2002. Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same? . *Journal of Environmental Economics and Management*, 44(3):391-425.

Carlson, J. L., 2000. Hypothetical Surveys Versus Real Economic Commitments: Further Evidence. *Applied Economics Letters*, 7(7):447-450.

Carlsson, F., Daruvala, D., and Jaldell, H., 2010. Do You Do What You Say or Do You Do What You Say Others Do? *Journal of Choice Modelling*, 3(2):113–133.

Carlsson, F., and Martinsson, P., 2001. Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? *Journal of Environmental Economics and Management*, 41(2):179-192.

Carson, K. S., Chilton, S. M., and Hutchinson, G. W., 2009. Necessary Conditions for Demand Revelation in Double Referenda. *Journal of Environmental Economics and Management*, 57(2):219–225.

Carson, R. T., and Czajkowski, M., 2014. The Discrete Choice Experiment Approach to Environmental Contingent Valuation. In: *Handbook of choice modelling*, S. Hess and A. Daly, eds., Edward Elgar, Northampton, MA.

Carson, R. T., Flores, N. E., Martin, K. M., and Wright, J. L., 1996. Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods. *Land Economics*, 72(1):80-99.

Carson, R. T., Flores, N. E., and Mitchell, R. C., 2001. Theory and Measurement of Passive-Use Value. In: *Valuing Environmental Preferences*, I. J. Bateman and K. G. Willis, eds., Oxford University Press Inc. , New York, 17-41.

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

Carson, R. T., Groves, T., and List, J. A., 2014. Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice. *Journal of the Association of Environmental and Resource Economists*, 1(1/2):171-207.

Carson, R. T., Hanemann, M., and Mitchell, R. C. (1987). "The Use of Simulated Political Markets to Value Public Goods." In: *Discussion Paper 87-7*, Department of Economics, Universityof California, San Diego.

Carson, R. T., and Louviere, J. J., 2011. A Common Nomenclature for Stated Preference Elicitation Approaches. *Environmental and Resource Economics*, 49(4):539-559.

Champ, P., and Bishop, R., 2001. Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias. *Environmental and Resource Economics*, 19(4):383-402.

Champ, P., Bishop, R., Brown, T., and McCollum, D., 1997. Using Donation Mechanisms to Value Non-use Benefits from Public Goods. *Journal of Environmental Economics and Management*, 33(2):151-162.

Champ, P., and Brown, T., 1997. A Comparison of Contingent and Actual Voting Behavior. In: *Proceedings from W-133 Benefits and Cost Transfer in Natural Resource Planning, 10th Interim Report*, 77-98.

Champ, P., Moore, R., and Bishop, R., 2009. A Comparison of Approaches to Mitigate Hypothetical Bias *Agricultural and Resource Economics Review*, 38(2):166-180.

Chang, J. B., Lusk, J. L., and Norwood, F. B., 2009. How Closely Do Hypothetical Surveys and Laboratory Experiments Predict Field Behavior? *American Journal of Agricultural Economics*, 91(2):518-534.

Choi, A. S., Ritchie, B. W., Papandrea, F., and Bennett, J., 2010. Economic Valuation of Cultural Heritage Sites: A Choice Modeling Approach. *Tourism Management*, 31(2):213-220.

Chowdhury, S., Meenakshi, J. V., Tomlins, K. I., and Owori, C., 2011. Are Consumers in Developing Countries Willing to Pay More for Micronutrient-Dense Biofortified Foods? Evidence from a Field Experiment in Uganda. *American Journal of Agricultural Economics*, 93(1):83-97.

Clarke, P. M., 2002. Testing the Convergent Validity of the Contingent Valuation and Travel Cost Methods in Valuing the Benefits of Health Care. *Health Economics*, 11(2):117–127.

Collins, J. P., and Vossler, C. A., 2009. Incentive compatibility tests of choice experiment value elicitation questions. *Journal of Environmental Economics and Management*, 58(2):226-235.

Cummings, R. G., Elliott, S., Harrison, G. W., and Murphy, J. J., 1997. Are Hypothetical Referenda Incentive Compatible? *Journal of Political Economy*, 105(3):609-621.

Cummings, R. G., Harrison, G. W., and Rutstrom, E. E., 1995. Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible? *American Economic Review*, 85(1):260-266.

Cummings, R. G., and Taylor, L. O., 1998. Does Realism Matter in Contingent Valuation Surveys? *Land Economics*, 74(2):203-215.

Cummings, R. G., and Taylor, L. O., 1999. Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method. *American Economic Review*, 89(3):649-665.

DeShazo, J. R., 2002. Designing Transactions Without Framing Effects in Iterative Question Formats. *Journal of Environmental Economics and Management*, 43:360-385.

Dickie, M., Fisher, A., and Gerking, S., 1987. Market Transactions and Hypothetical Demand Data: A Comparative Study. *Journal of the American Statistical Association*, 82(397):69-75.

Duffield, J. W., and Patterson, D. A. (1992). "Field Testing Existence Values: Comparison of Hypothetical and Cash Transaction Values." In: *Joint Western Regional Science Association Session on Measuring Option and Existence Values*, South Lake Tahoe, Nevada.

Ethier, R. G., Poe, G. L., Schulze, W., and Clark, J. E., 2000. A Comparison of Hypothetical Phone and Mail Contingent Valuation Responses for Green-Pricing Electricity Programs. *Land Economics*, 76(1):54-67.

Farquharson, R., 1969. Theory of Voting. Yale University Press, New Haven.

Ferrini, S., Schaafsma, M., and Bateman, I. J., 2014. Revealed and Stated Preference Valuation and Transfer: A Within-Sample Comparison of Water Quality Improvement Values. *Water Resources Research*, 50(6):4746–4759.

Fifer, S., Rose, J., and Greaves, S., 2014. Hypothetical Bias in Stated Choice Experiments: Is It a Problem? And If So, How Do We Deal With It? *Transportation Research Part A: Policy and Practice*, 61:164-177.

Foster, V., Bateman, I. J., and Harley, D., 1997. Real and Hypothetical Willingness to Pay for Environmental Preservation: A Non-Experimental Comparison. *Journal of Agricultural Economics*, 48(2):123-138.

Frykblom, P., 1997. Hypothetical Question Modes and Real Willingness to Pay. *Journal of Environmental Economics and Management*, 34(3):275-287.

Getzner, M., 2000. Hypothetical and Real Economic Commitments, and Social Status, in Valuing a Species Protection Programme. *Journal of Environmental Planning and Management*, 43(4):541-559.

Gibbard, A., 1973. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41:587–601.

Hensher, D. A., Rose, J. M., and Greene, W. H., 2005. Applied Choice Analysis: A Primer. Cambridge University Press, Cambridge.

Herriges, J., Kling, C., Liu, C.-C., and Tobias, J., 2010. What are the consequences of consequentiality? *Journal of Environmental Economics and Management*, 59(1):67-81.

Herriges, J. A., and Shogren, J. F., 1996. Starting Point Bias in Dichotomous Choice Valuation with Follow-up Questioning. *Journal of Environmental Economics and Management*, 30(1):112-131.

Hoehn, J. P., and Randall, A., 1987. A Satisfactory Benefit Cost Indicator from Contingent Valuation. *Journal of Environmental Economics and Management*, 14(3):226-247.

Hudson, D., Gallardo, R. K., and Hanson, T. R., 2012. A Comparison of Choice Experiments and Actual Grocery Store Behavior: An Empirical Application to Seafood Products. *Journal of Agricultural and Applied Economics*, 44(1):49-62.

Hwang, J., Petrolia, D. R., and Interis, M. G., 2014. Valuation, Consequentiality, and Opt-Out Responses to Stated Preference Surveys. *Agricultural and Resource Economics Review*, 43(3):471–488.

Isacsson, G. (2007). "The Trade Off Between Time and Money: Is There a Difference Between Real and Hypothetical Choices?" In: *Working Papers no. 2007:3*, Swedish National Road & Transport Research Institut.

Johannesson, M., 1997. Some Further Experimental Results on Hypothetical Versus Real Willingness to Pay. *Applied Economics Letters*, 4(8):535-536.

Johannesson, M., Liljas, B., and Johansson, P.-O., 1998. An Experimental Comparison of Dichotomous Choice Contingent Valuation Questions and Real Purchase Decisions. *Applied Economics*, 30(5):643-647.

Johansson-Stenman, O., and Svedsäter, H., 2008. Measuring Hypothetical Bias in Choice Experiments: The Importance of Cognitive Consistency. *The B.E. Journal of Economic Analysis & Policy*, 8(1):1-10.

Johnston, R., Ranson, M., Besedin, E., and Helm, E., 2006. What Determines Willingness to Pay per Fish? A Meta-Analysis of Recreational Fishing Values. *Marine Resource Economics*, 21(1):1-32.

Johnston, R. J., 2006. Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum. *Journal of Environmental Economics and Management*, 52(1):469-481.

Kahneman, D., and Knetsch, J. L., 1992. Valuing Public Goods: The Purchase of Moral Satisfaction. *Journal of Environmental Economics and Management*, 22:57-70.

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

Kealy, M. J., Dovidio, J. F., and Rockel, M. L., 1988. Accuracy in Valuation is a Matter of Degree. *Land Economics*, 64(2):158-171.

Kesternich, I., Heiss, F., McFadden, D., and Winter, J., 2013. Suit the Action to the Word, the Word to the Action: Hypothetical Choices and Real Decisions in Medicare Part D. *Journal of Health Economics*, 32(6):1313-1324.

Kling, C. L., Phaneuf, D. J., and Zhao, J., 2012. From Exxon to BP: Has Some Number Become Better than No Number? *The Journal of Economic Perspectives*, 26(4):3-26.

Knetsch, J. L., and Davis, R. K., 1966. Comparisons of Methods for Resource Evaluation. In: *Water Research*, A. V. Kneese and S. C. Smith, eds., Johns Hopkins Press for Resources for the Future, Baltimore

Kochi, I., Hubbell, B., and Kramer, R., 2006. An Empirical Bayes Approach to Combining and Comparing Estimates of the Value of a Statistical Life for Environmental Policy Analysis. *Environmental and Resource Economics*, 34(3):385-406.

Krawczyk, M., 2012. Testing for hypothetical bias in willingness to support a reforestation program. *Journal of Forest Economics*, 18(4):282-289.

Kurz, M., 1974. Experimental Approach to the Determination of Demand for Public Goods. *Journal of Public Economics*, 3(4):329-348.

Landry, C. E., and List, J. A., 2007. Using Ex Ante Approaches to Obtain Credible Signals for Value in Contingent Markets: Evidence from the Field. *American Journal of Agricultural Economics*, 89(2):420-429.

List, J. A., 2001. Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards. *American Economic Review*, 91(5):1498-1507.

List, J. A., and Gallet, C. A., 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics*, 20(3):241-254.

List, J. A., and Shogren, J. F., 1998. Calibration of The Difference Between Actual and Hypothetical Valuations in a Field Experiment. *Journal of Economic Behavior and Organization*, 37(2):193-206.

List, J. A., Sinha, P., and Taylor, M. H., 2006. Using Choice Experiments to Value Non-Market Goods and Services: Evidence from Field Experiments. *Advances in Economic Analysis and Policy*, 6(2):1-37.

Little, J., and Berrens, R., 2004. Explaining Disparities Between Actual and Hypothetical Stated Values: Further Investigation Using Meta-Analysis. *Economics Bulletin*, 3(6):1-13.

Loomis, J. B., Bell, P., Cooney, H., and Asmus, C., 2009. A Comparison of Actual and Hypothetical Willingness to Pay of Parents and Non-Parents for Protecting Infant Health: The Case of Nitrates in Drinking Water. *Journal of Agricultural and Applied Economics*, 41(3):697-712.

Loomis, J. B., Brown, T., Lucero, B., and Peterson, G., 1997. Evaluating the Validity of the Dichotomous Choice Question Format in Contingent Valuation. *Environmental and Resource Economics*, 10(2):109-123.

Loomis, J. B., and Gonzalez-Caban, A., 1997. Comparing the Economic Value of Reducing Fire Risk to Spotted Owl Habitat in California and Oregon. *Forest Science*, 43(4):473-482.

Loomis, J. B., Pierce, C., and Manfredo, M., 2000. Using the Demand for Hunting Lcences to Evaluate Contingent Valuation Estimates of Willingness to Pay. *Applied Economics Letters*, 7(7):435-438.

Louviere, J. J., Hensher, D. A., and Swait, J. D., 2006. Stated Choice Methods: Analysis and Applications. Cambridge University Press, Cambridge.

Lusk, J. L., Pruitt, J. R., and Norwood, F. B., 2006. External Validity of a Framed Field Experiment. *Economics Letters*, 93(2):285-290.

Lusk, J. L., and Schroeder, T. C., 2004. Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks. *American Journal of Agricultural Economics*, 86(2):467-482.

MacMillan, D., Smart, T., and Thorburn, A., 1999. A Field Experiment Involving Cash and Hypothetical Charitable Donations. *Environmental and Resource Economics*, 14(3):399-412.

Mitani, Y., and Flores, N. E., 2009. Demand Revelation, Hypothetical Bias, and Threshold Public Goods Provision. *Environmental and Resource Economics*, 44(2):231-243.

Mitani, Y., and Flores, N. E. (2012). "Robustness Tests of Incentive Compatible Referenda: Consequential Probability, Group Size, and Value-cost Difference." In: *European Association of Environmental and Resource Economists 19th Annual Conference, June 27 - 30*, Prague, Czech Rebuplic.

Mitchell, R. C., and Carson, R. T., 1989. Using Surveys to Value Public Goods: The Contingent Valuation Methods. Resources for the Future, Washington.

Moser, R., Raffaelli, R., and Notaro, S., 2014. Testing Hypothetical Bias With a Real Choice Experiment Using Respondents' Own Money. *European Review of Agricultural Economics*, 41(1):25-46.

Mozumder, P., and Berrens, R., 2007. Investigating Hypothetical Bias: Induced-Value Tests of the Referendum Voting Mechanism with Uncertainty. *Applied Economics Letters*, 14(10):705-709.

Murphy, J. J., Allen, P. G., Stevens, T., and Weatherhead, D., 2005a. A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics*, 30(3):313-325.

Murphy, J. J., Stevens, T., and Weatherhead, D., 2005b. Is Cheap Talk Effective at Eliminating Hypothetical Bias in a Provision Point Mechanism? *Environmental and Resource Economics*, 30(3):327-343.

Murphy, J. J., Stevens, T., and Yadav, L., 2010. A Comparison of Induced Value and Home-Grown Value Experiments to Test for Hypothetical Bias in Contingent Valuation. *Environmental and Resource Economics*, 47(1):111-123.

Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W., and McGuckin, T., 1994. Hypothetical Surveys and Real Economic Commitments. *Land Economics*, 70(2):145-154.

Nepal, M., Berrens, R., and Bohara, A. K., 2009. Assessing Perceived Consequentiality: Evidence From a Contingent Valuation Survey on Global Climate Change. *International Journal of Ecological Economics and Statistics*, 14(P09):14-29.

Nocera, S., Telser, H., and Bonato, D., 2003. The Contingent Valuation Method in Health Care. Springer.

Norwood, F. B., and Lusk, J. L., 2011. Social Desirability Bias in Real, Hypothetical, and Inferred Valuation Experiments. *American Journal of Agricultural Economics*, 93(2):528-534.

Paradiso, M., and Trisorio, A., 2001. The Effect of Knowledge on the Disparity between Hypothetical and Real Willingness to Pay. *Applied Economics*, 33(11):1359-1364.

Poe, G. L., Clark, J. E., Rondeau, D., and Schulze, W., 2002. Provision Point Mechanisms and Field Validity Tests of Contingent Valuation. *Environmental and Resource Economics*, 23(1):105-131.

Polomé, P., 2003. Experimental Evidence on Deliberate Misrepresentation in Referendum Contingent Valuation. *Journal of Economic Behavior and Organization*, 52(3):387-401.

Rasmusen, E., 1989. Games and Information: An Introduction to Game Theory. Blackwell, New York.

Ready, R. C., Champ, P., and Lawton, J. L., 2010. Using Respondent Uncertainty to Mitigate Hypothetical Bias in a Stated Choice Experiment. *Land Economics*, 86(2):363-381.

Rolfe, J., and Dyack, B., 2010. Testing for Convergent Validity Between Travel Cost and Contingent Valuation Estimates of Recreation Values in the Coorong, Australia. *Australian Journal of Agricultural and Resource Economics*, 54(4):583–599.

Rosenberger, R. S., and Loomis, J. B., 2000. Using Meta-analysis for Benefit Transfer: In-sample Convergent Validity Tests of an Outdoor Recreation Database. *Water Resources Research*, 36(4):1097-1107.

Satterthwaite, M. A., 1975. Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems of Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, 10(2):187–217.

Schläpfer, F., Roschewitz, A., and Hanley, N., 2004. Validation of Stated Preferences for Public Goods: A Comparison of Contingent Valuation Survey Response and Voting Behaviour. *Ecological Economics*, 52(1-2):1-16.

Seip, K., and Strand, J., 1992. Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment. *Envwomnental and Resource Economics*, 2(1):91-106.

Shogren, J. F., Fox, J. A., Hayes, D. J., and Roosen, J., 1999. Observed for Food Safety in Retail, Survey, and Auction Markets. *American Journal of Agricultural Economics*, 81(5):1192-1199.

Shrestha, R. K., and Loomis, J. B., 2003. Meta-Analytic Benefit Transfer of Outdoor Recreation Economic Values: Testing Out-of-Sample Convergent Validity. *Environmental and Resource Economics*, 25(1):79–100.

Sinden, J. A., 1988. Empirical Tests of Hypothetical Bias in Consumer's Surplus Surveys. *Australian Journal of Agricultural Economics*, 32(2-3):98-112.

Smith, V. K., and Mansfield, C., 1998. Buying Time: Real and Hypothetical Offers. *Journal of Environmental Economics and Management*, 36(3):209-224.

Spencer, M. A., Swallow, S. K., and Miller, C. J., 1998. Valuing Water Quality Monitoring: A Contingent Valuation Experiment Involving Hypothetical and Real Payments. *Agricultural and Resource Economics Review*, 27(1):28-42.

Stefani, G., and Scarpa, R. (2009). "The Referendum Incentive Compatibility Hypothesis: Some New Results Using Information Messages." In: *Working Papers in Economics no. 07/10*, University of Waikato, Department of Economics, New Zeland.

Swallow, B. M., and Woudyalew, M., 1994. Evaluating Willingness to Contribute to a Local Public Good: Application of Contingent Valuation to Tsetse Control in Ethiopia. *Ecological Economics*, 11(2):153-161.

Taylor, L. O., 1998. Incentive Compatible Referenda and the Valuation of Environmental Goods. *Agricultural and Resource Economics Review*, 27(2):132-139.

Taylor, L. O., McKee, M., Laury, S. K., and Cummings, R. G., 2001. Induced-Value Tests of the Referendum Voting Mechanism. *Economics Letters*, 71(1):61-65.

Taylor, L. O., Morrison, M. D., and Boyle, K. J., 2010. Exchange Rules and the Incentive Compatibility of Choice Experiments. *Environmental and Resource Economics*, 47(2):197-220.

Veisten, K., and Navrud, S., 2006. Contingent Valuation and Actual Payment for Voluntarily Provided Passive-Use Values: Assessing the Effect of an Induced Truth-Telling Mechanism and Elicitation Formats. *Applied Economics*, 38(7):735-756.

Volinskiy, D., Adamowicz, W., and Veeman, M., 2011. Predicting Versus Testing: A Conditional Cross-Forecasting Accuracy Measure for Hypothetical Bias. *The Australian Journal of Agricultural and Resource Economics*, 55(3):429-450.

Vossler, C. A., Doyon, M., and Rondeau, D., 2012a. Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics*, 4(4):145-171.

Vossler, C. A., Doyon, M., Rondeau, D., and Roy-Vigneault, F., 2012b. Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics*, 4(4):145-171.

Vossler, C. A., and Evans, M. F., 2009. Bridging the gap between the field and the lab: Environmental goods, policy maker input, and consequentiality. *Journal of Environmental Economics and Management*, 58(3):338-345.

Vossler, C. A., and Kerkvliet, J., 2003. A Criterion Validity Test of the Contingent Valuation Method: Comparing Hypothetical and Actual Voting Behavior for a Public Referendum. *Journal of Environmental Economics and Management*, 45(3):631-649.

Vossler, C. A., Kerkvliet, J., Polasky, S., and Gainutdinova, O., 2003. Externally Validating Contingent Valuation: An Open-Space Survey and Referendum in Corvallis, Oregon. *Journal of Economic Behavior and Organization*, 51(2):261-277.

Vossler, C. A., and McKee, M., 2006. Induced-Value Tests of Contingent Valuation Elicitation Mechanisms. *Environmental and Resource Economics*, 35(2):137-168.

Vossler, C. A., and Watson, S. B., 2013. Understanding the Consequences of Consequentiality: Testing the Validity of Stated Preferences in the Field. *Journal of Economic Behavior and Organization*, 86:137-147.

Walsh, R. G., Johnson, D. M., and McKean, J. R., 1989. Issues in Nonmarket Valuation and Policy Application: A Retrospective Glance. *Western Journal of Agricultural Economics*, 14(1):178–188.

Walsh, R. G., Johnson, D. M., and McKean, J. R., 1992. Benefit Transfer of Outdoor Recreation Demand Studies: 1968-1988. *Water Resources Research*, 28(3):707–713.

Whitehead, J. C., Phaneuf, D. J., Dumas, C. F., Herstine, J., Hill, J., and Buerger, B., 2010. Convergent Validity of Revealed and Stated Recreation Behavior with Quality Change: A Comparison of Multiple and Single Site Demands. *Environmental and Resource Economics*, 45(1):91-112.

Woodward, R. T., and Wui, Y. S., 2000. The Economic Value of Wetland Services: A Meta-analysis. *Ecological Economics*, 37(2):257–270.

Yue, C., and Tong, C., 2009. Organic or Local? Investigating Consumer Preference for Fresh Produce Using a Choice Experiment with Real Economic Incentives. *HortScience*, 44(2):366-371.