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

EWA ZAWOJSKA
Faculty of Economic Sciences
University of Warsaw
e-mail: ezawojska@wne.uw.edu.pl

MIKOŁAJ CZAJKOWSKI
Faculty of Economic Sciences
University of Warsaw
e-mail: miq@wne.uw.edu.pl

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 considered, 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.

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

JEL:
Q51, H4, D6

Acknowledgements
The second author gratefully acknowledges the support of the Polish Ministry of Science and Higher Education and the Foundation for Polish Science.
1. Introduction

Stated preference (SP) data is commonly collected in surveys to enable researchers to model consumers' preferences and thus determine valuations for the goods or policies under investigation. 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 valuable because 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, Murphy et al. 2005, Little and Berrens 2004).

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 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) consequentiaility, 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.

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

---

1 Applications of survey-based methods to determine economic preferences that are common not only in environmental economics (Kanninen 2007) but also in marketing (Louviere, Hensher, and Swait 2006), transport (Hensher, Rose, and Greene 2005), health (Nocera, Telser, and Bonato 2003), culture (Choi et al. 2010), and many other fields.

2 Incentive compatibility means that a respondent’s optimal strategy is to answer the survey truthfully.
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 individuals’ 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’ preferences expressed in surveys constructed in a particular manner. A typical SP 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 calculating nonmarket valuations. 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, Flores, and Mitchell 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 matching methods or discrete choice experiments (DCE) (Carson and Louviere 2011). In matching methods, respondents indicate a specific number that usually expresses their WTP. Direct questions 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 wide range of monetary values, from which they choose the one representing their maximum WTP. In contrast to matching methods, DCE surveys typically ask participants 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

<table>
<thead>
<tr>
<th>Number of choice tasks (CT)</th>
<th>Number of choice alternatives (A)</th>
<th>Number of choice alternatives (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A = 2$</td>
<td>$A &gt; 2$</td>
</tr>
<tr>
<td>$CT = 1$</td>
<td>Single binary choice</td>
<td>Single multinomial choice</td>
</tr>
<tr>
<td>$CT &gt; 1$</td>
<td>Binary choice sequence</td>
<td>Multinomial choice sequence</td>
</tr>
</tbody>
</table>

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 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 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 also 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 results obtained from SP data with those from RP data.

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 but typically elicits consumers’ preferences for the same or very similar good in both hypothetical and actual payment settings. The real-payment-based estimates then provide a reference point for validity verification.

The consistency between WTP estimates obtained from SP studies and RP or actual payment conditions 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,” as it is usually ascribed to the hypothetical nature of SP surveys, which provide different incentives than those experienced in real-life situations. The issue of the incentive compatibility 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.³

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 multinominal choice question, if a respondent is convinced that the agency will introduce only one of the proposed options, then 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

³ 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 nonetheless 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).
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 independently consider all choice tasks in a sequence. 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.4

In summary, to elicit accurate preferences, an 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 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

4 The payment card mechanism, which was designed to overcome problems tied to the direct question, also fails to offer incentive-compatible conditions. This format can be viewed as a form of 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.
investigated, 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 researches 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, 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, Johnson, and McKean (1989) and Walsh, Johnson, and McKean (1992) have conducted the first meta-analyses of recreation valuation studies to assess the discrepancy between estimates from CV and travel cost methods. They find that RP estimates usually

---

5 A noteworthy exception is the requirement of consequentiality, as theoretical and empirical evidence of its importance continues to mount (Herriges et al. 2010, Nepal, Berrens, and Bohara 2009, Vossler, Doyon, and Rondeau 2012, Vossler and Watson 2013, Vossler and Evans 2009, Broadbent 2012).

6 The weighted dataset treats the mean SP to RP ratio from each study as one observation when the study provides multiple estimates.
exceed those derived using 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. 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, Schaafsma, and Bateman (2014) have also reported contradictory evidence, i.e., travel cost and CV payment card estimates do not differ significantly, whereas dichotomous choice values 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 from 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 also depends on other characteristics such as the year a study was conducted.

Brander, Van Beukering, and Cesar (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.

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 travel cost methods.7 Brander, Florax, and Vermaat (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 techniques, including hedonic pricing and travel cost methods, which essentially contradicts earlier findings.

7 These results should be exercised with caution because the sample included only two studies that applied the hedonic price method.
Aadland and Caplan (2003) examine studies on curbside recycling valuation and find higher WTP estimates based on SP studies compared to those from RP studies. They suggest that cheap talk\(^8\) scripts could be an efficient method to attenuate this discrepancy.

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, Rose, and Greaves (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, Hubbell, and Kramer (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

\(^8\) 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.
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. Unfortunately, such a comparison is currently unavailable. Its usefulness would, however, still be 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.

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 from actual and simulated market studies, induced-value lab experiments, or naturally occurring public referenda. In what follows, we review the empirical evidence from criterion validity tests, which we have grouped into these four categories.

3.2.1. Actual market studies
Convergent 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 so that there are markets in which purchasing behavior can be observed. 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, Heberlein, and Kealy (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 actual market behavior. A similar relationship is reported by Loomis, Pierce, and Manfredo (2000). 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, Chowdhury et al. 2011, List and Shogren 1998, Loomis et al. 2009) 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.

---

9 Another point raised in the support of actual market studies is respondents’ familiarity with the good. This, however, is not required by any of the incentive compatibility conditions.
Obtaining different WTP estimates in SP surveys compared to actual market studies is not a rule. For example, Dickie, Fisher, and Gerking (1987), Smith and Mansfield (1998), List, Sinha, and Taylor (2006), and Lusk, Pruitt, and Norwood (2006) 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.

Most studies reported above are not consequential, which alone offers sufficient grounds to question their conclusions. Carson, Groves, and List (2014a) and Landry and List (2007) compare SP and actual market choices in a consequential setting. Carson, Groves, and List (2014a) use various levels of the probability of a survey being binding, which are ascribed to different treatments. They 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.

Another stream of criterion validity literature uses donations as a payment mechanism, typically asking respondents about their willingness to contribute to the provision of a public good (for example, Brown and Duffield 1995, Loomis and Gonzalez-Caban 1997, Swallow and Woudyalew 1994, Seip and Strand 1992b). After all, as argued by Champ et al. (1997), respondents may see donations as a more plausible result from a survey outcome than a tax increase. The results of these studies, summarized in the Appendix Table A2, show that respondents’ actual voluntary contributions are often overestimated by SP questions.

Although it might seem to be evidence contesting the validity of SP methods, higher declared contributions compared to actual contributions are in fact expected by the economic theory. When discussing real donations, rational participants free ride and let the public good be provided through contributions from others, although it makes sense to state a WTP in order
to have the good be seen as worth providing. In addition, a contingent scenario that does not include a coercive payment mechanism is known *not* to be incentive compatible because it encourages respondents to overstate their WTP, 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.

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 (Carson, Groves, and List 2014b, Landry and List 2007, Herriges et al. 2010) 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 SP methods’ 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, quasi-public), employing various mechanisms to determine the final outcome (donations, referenda), and applying either home-grown or induced values.10

---

10 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.
Numerous criterion validity tests in a laboratory setting have been conducted using private goods (Appendix Table A3 provides an overview). Again, the evidence resulting from these studies is mixed and does not allow univocal conclusions to be drawn. Some researchers observe a statistically significant discrepancy between estimates from hypothetical and actual payment conditions, whereas others document that hypothetical CVs do not produce significantly different values. Nevertheless, we argue that since the essential prerequisite for truthful preference revelation is adherence to incentive compatibility conditions, most of the available studies do not provide a valuable input to the discussion because they rarely utilize a single binary choice question. Moreover, 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. 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 A4. Most show a significant divergence between respondents’ behavior in 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 studies applying the donation mechanism, only Broadbent, Grandy, and Berrens (2010) and Broadbent (2012) consider the role of consequentiality. They find that respondents

---

11 A possible exception is the experiment reported by Johannesson, Liljas, and Johansson (1998), who emphasize that respondents should provide the amount they would pay “here and now” in contrast to the usual “would you ever pay...” In conjunction with an incentive-compatible elicitation format, the results show no significant discrepancy between values resulting from a CV survey and an actual payment treatment.

12 Herriges et al. (2010) is a notable example of a study that simulates an out-of-laboratory, public good market using incentive-compatible elicitation question format and could have been perceived by respondents as consequential. They 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. They provide evidence to the SP method’s validity, as long as incentive compatibility conditions are satisfied.
believing in survey consequentiality\textsuperscript{13} 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.\textsuperscript{14} 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 and complying with the consequentiality requirement\textsuperscript{15} are reported by Cummings and Taylor (1998), Vossler and Evans (2009), and Vossler et al. (2012). 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 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

\textsuperscript{13} 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.

\textsuperscript{14} In contrast, Herriges et al. (2010) measure consequentiality perception through self-reports on a five-degree Likert scale.

\textsuperscript{15} 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 tax increase, which excludes the possibility of free riding.
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. (2012) 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 A5 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, the conclusion becomes even stronger, i.e., no significant discrepancy between stated and true values exists.

---

16 Carson, Chilton, and Hutchinson (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.
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 (Bateman et al. 2002, Poe et al. 2002, Taylor 1998). 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 A6 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, Hanemann, and Mitchell (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 survey and actual votes. Schläpfer, Roschewitz, and Hanley (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. Field experiments could shed more light both on their validity in general and on whether lab tests of CV indeed provide valuable input. Second, research on consequentiality perception is still in its infancy (Kling, Phaneuf, and Zhao 2012). Crucial questions concern measuring the perceived level of consequentiality (Nepal, Berrens, and Bohara 2009, Herriges et al. 2010) and making the measurements themselves incentive compatible. Finally, there are currently no studies providing an indication of the extent 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 has accrued concerning the implementation of various SP methods, and much has been learned. One of the main areas where SP methods 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 this condition is considered, SP methods appear to consistently provide valid results. This result is reassuring and indicates that if designed and conducted appropriately, SP methods will remain of central importance to modern welfare economics and environmental economics in particular.

---

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


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

<table>
<thead>
<tr>
<th>Author</th>
<th>Good</th>
<th>Elicitation mode</th>
<th>Elicitation format(^{18})</th>
<th>Sample</th>
<th>Divergence(^{19})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bishop and Heberlein (1979)</td>
<td>goose hunting permits</td>
<td>mail</td>
<td>SBC followed by DQ</td>
<td>split</td>
<td>significant, downward</td>
</tr>
<tr>
<td>Bishop, Heberlein, and Kealy (1983)</td>
<td>goose hunting permits</td>
<td>mail</td>
<td>SBC followed by DQ</td>
<td>split</td>
<td>significant, downward</td>
</tr>
<tr>
<td>Carson, Groves, and List (2014a)</td>
<td>baseball memorabilia</td>
<td>field referendum</td>
<td>SBC</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Chang, Lusk, and Norwood (2009)</td>
<td>dishwashing liquid, ground beef, wheat flour</td>
<td>field survey, real store</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Chowdhury et al. (2011)</td>
<td>biofortified staple crops</td>
<td>field survey</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Dickie, Fisher, and Gerking (1987)</td>
<td>fresh strawberries</td>
<td>in-person interview</td>
<td>DQ</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Hudson, Gallardo, and Hanson (2012)</td>
<td>seafood products</td>
<td>mail, real store</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Landry and List (2007)</td>
<td>sports memorabilia</td>
<td>field referendum</td>
<td>B-SEQ (double) DQ (CV); second price bid auction (actual)</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>List and Shogren (1998)</td>
<td>baseball cards</td>
<td>field survey</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Isacsson (2007)</td>
<td>bus journey</td>
<td>field survey</td>
<td>SBC</td>
<td>split</td>
<td>significant, downward</td>
</tr>
<tr>
<td>List, Sinha, and Taylor (2006)</td>
<td>sports cards</td>
<td>field survey</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Loomis et al. (2009)</td>
<td>bottled water protecting infant health</td>
<td>in-person interview, mail</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Loomis, Pierce, and Manfredo (2000)</td>
<td>elk and deer hunting permits</td>
<td>phone</td>
<td>SBC (CV); variations in prices (actual)</td>
<td>split</td>
<td>significant, downward</td>
</tr>
<tr>
<td>Lusk, Pruitt, and Norwood (2006)</td>
<td>pork chops</td>
<td>in-person interview, real store</td>
<td>SMC</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Moser, Raffaelli, and Notaro (2014)</td>
<td>apples</td>
<td>in-person interview</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Shogren et al. (1999)</td>
<td>chicken breasts</td>
<td>mail, real store</td>
<td>SBC</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Smith and Mansfield (1998)</td>
<td>opportunity cost of survey participation</td>
<td>phone</td>
<td>SBC</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Yue and Tong (2009)</td>
<td>organic and locally grown tomatoes</td>
<td>field survey</td>
<td>M-SEQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
</tbody>
</table>

---

\(^{18}\) The following notation is used (applies to all tables in the Appendix):
- SBQ – a single binary question,
- B-SEQ – a binary choice sequence,
- SMC – a single multinomial choice question,
- M-SEQ – a multinomial choice sequence,
- DQ – a direct question,
- PC – a payment card.

\(^{19}\) “Divergence” expresses the divergence of CV-based estimates from actual values.
Appendix A2. Criterion validity tests in actual market studies using donations

<table>
<thead>
<tr>
<th>Author</th>
<th>Elicitation mode</th>
<th>Elicitation format</th>
<th>Sample</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown et al. (1996)</td>
<td>mail</td>
<td>SBC, DQ</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Brown and Taylor (2000)</td>
<td>in-person interview</td>
<td>DQ</td>
<td></td>
<td>significant, upward</td>
</tr>
<tr>
<td>Cameron et al. (2002)</td>
<td>mail (CV), phone (actual)</td>
<td>SBC, DQ, PC</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Champ and Bishop (2001)</td>
<td>mail</td>
<td>SBC, DB</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Champ et al. (1997)</td>
<td>mail</td>
<td>SBC</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Duffield and Patterson (1992)</td>
<td>phone, mail</td>
<td>SBC</td>
<td></td>
<td>significant, upward</td>
</tr>
<tr>
<td>Ethier et al. (2000)</td>
<td>mail</td>
<td>DQ</td>
<td></td>
<td>significant, upward</td>
</tr>
<tr>
<td>Foster, Bateman, and Harley (1997)</td>
<td>phone, mail</td>
<td>SBC</td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>List, Sinha, and Taylor (2006)</td>
<td>field</td>
<td>SMC</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>MacMillan, Smart, and Thorburn (1999)</td>
<td>mail</td>
<td>DQ</td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>Poe et al. (2002)</td>
<td>phone</td>
<td>DB</td>
<td></td>
<td>significant, upward</td>
</tr>
<tr>
<td>Seip and Strand (1992a)</td>
<td>mail (CV), phone (actual)</td>
<td>SBC</td>
<td>within</td>
<td>significant, upward</td>
</tr>
</tbody>
</table>
### Appendix A3. Criterion validity tests in simulated private good market studies

<table>
<thead>
<tr>
<th>Author</th>
<th>Elicitation format</th>
<th>Sample</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfnes et al. (2006)</td>
<td>M-SEQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Aoki, Shen, and Tatsuyoshi (2010)</td>
<td>M-SEQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Balistreri et al. (2001)</td>
<td>DQ, SBC</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Camacho-Cuena et al. (2004)</td>
<td>PC</td>
<td>insignificant, upward</td>
<td></td>
</tr>
<tr>
<td>Carlson (2000)</td>
<td>SBC followed by DQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Cummings, Harrison, and Rutstrom (1995)</td>
<td>SBC</td>
<td>significant</td>
<td></td>
</tr>
<tr>
<td>Frykblom (1997)</td>
<td>SBC, DQ</td>
<td>insignificant</td>
<td></td>
</tr>
<tr>
<td>Johannesson (1997)</td>
<td>DQ (CV), second-price auction (actual)</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Johanesson, Liljas, and Johannsson (1998)</td>
<td>SBC</td>
<td>insignificant</td>
<td></td>
</tr>
<tr>
<td>Kealy, Dovidio, and Rockel (1988)</td>
<td>SBC followed by DQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Loomis et al. (1997)</td>
<td>SBC, DQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Lusk and Schroeder (2004)</td>
<td>M-SEQ</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Neill et al. (1994)</td>
<td>DQ (CV), Vickery auction (CV and actual)</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Paradiso and Trisorio (2001)</td>
<td>DQ (CV), Vickery auction (actual)</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Stefani and Scarpa (2009)</td>
<td>SBC</td>
<td>significant, upward</td>
<td></td>
</tr>
<tr>
<td>Taylor, Morrison, and Boyle (2010)</td>
<td>M-SEQ</td>
<td>insignificant</td>
<td></td>
</tr>
<tr>
<td>Volinskiy, Adamowicz, and Veeman (2011)</td>
<td>B-SEQ, M-SEQ</td>
<td>split, significant</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix A4. Criterion validity tests in simulated public good market studies

<table>
<thead>
<tr>
<th>Author</th>
<th>Elicitation format</th>
<th>Choice setting</th>
<th>Sample</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botelho and Pinto (2002)</td>
<td>OE</td>
<td>Donation</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Broadbent (2012)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Broadbent (2014)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Broadbent, Grandy, and Berrens (2010)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Carlsson, Daruvala, and Jaldell (2010)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split</td>
<td>significant</td>
</tr>
<tr>
<td>Carlsson and Martinsson (2001)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>within</td>
<td>insignificant</td>
</tr>
<tr>
<td>Getzner (2000)</td>
<td>DQ and SBC</td>
<td>Donation</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Johansson-Stenman and Svedsäter (2008)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Murphy, Stevens, and Weatherhead (2005)</td>
<td>SBC</td>
<td>Donation</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Ready, Champ, and Lawton (2010)</td>
<td>M-SEQ</td>
<td>Donation</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Cummings et al. (1997)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Cummings and Taylor (1998)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Cummings and Taylor (1999)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Krawczyk (2012)</td>
<td>B-SEQ</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Stefani and Scarpa (2009)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Taylor (1998)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Vossler et al. (2012)</td>
<td>B-SEQ</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Vossler and Evans (2009)</td>
<td>SBC</td>
<td>Referendum</td>
<td>split, within</td>
<td>significant, upward</td>
</tr>
</tbody>
</table>


## Appendix A5. Criterion validity tests using induced-value experiments

<table>
<thead>
<tr>
<th>Author</th>
<th>Elicitation format</th>
<th>Choice setting</th>
<th>Sample</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burton et al. (2007)</td>
<td>SBC</td>
<td>Other</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Carson, Chilton, and Hutchinson (2009)</td>
<td>B-SEQ</td>
<td>Referendum</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Collins and Vossler (2009)</td>
<td>B-SEQ, M-SEQ</td>
<td>Other</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Mitani and Flores (2009)</td>
<td>OE</td>
<td>Donation</td>
<td>within</td>
<td>insignificant</td>
</tr>
<tr>
<td>Mitani and Flores (2012)</td>
<td>B-SEQ</td>
<td>Other</td>
<td>within</td>
<td>insignificant</td>
</tr>
<tr>
<td>Mozumder and Berrens (2007)</td>
<td>B-SEQ</td>
<td>Other</td>
<td>split</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Murphy, Stevens, and Yadav (2010)</td>
<td>B-SEQ</td>
<td>Other</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Polomé (2003)</td>
<td>B-SEQ</td>
<td>Other</td>
<td>---</td>
<td>insignificant</td>
</tr>
<tr>
<td>Taylor et al. (2001)</td>
<td>SBC</td>
<td>Other</td>
<td>split</td>
<td>insignificant</td>
</tr>
<tr>
<td>Vossler and McKee (2006)</td>
<td>SBC, PC, multiple-bounded</td>
<td>Other</td>
<td>split</td>
<td>insignificant</td>
</tr>
</tbody>
</table>
### Appendix A6. Criterion validity tests based on naturally occurring referenda

<table>
<thead>
<tr>
<th>Author</th>
<th>Elicitation format</th>
<th>Sample</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carson, Hanemann, and Mitchell (1987)</td>
<td></td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>Champ and Brown (1997)</td>
<td></td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>Johnston (2006)</td>
<td></td>
<td>SBC</td>
<td>insignificant</td>
</tr>
<tr>
<td>Schläpfer, Roschewitz, and Hanley (2004)</td>
<td></td>
<td>B-SEQ</td>
<td>significant, upward</td>
</tr>
<tr>
<td>Vossler and Kerkvliet (2003)</td>
<td></td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>Vossler et al. (2003)</td>
<td></td>
<td></td>
<td>insignificant</td>
</tr>
<tr>
<td>Vossler and Watson (2013)</td>
<td></td>
<td></td>
<td>insignificant</td>
</tr>
</tbody>
</table>