Can cheap panel-based internet surveys substitute costly in-person interviews in CV surveys?

Henrik Lindhjem and Ståle Navrud

Norwegian University of Life Sciences, Norwegian Institute for Nature Research (NINA)

9. July 2010

Can cheap panel-based internet surveys substitute costly in-person interviews in CV surveys?

Henrik Lindhjem[a,1], Ståle Navrud[a],

[a] Department of Economics and Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway.

Date of draft: 9. July 2010

[1] Corresponding author: Norwegian Institute for Nature Research (NINA), Gaustadalleen 21, N-0349 Oslo, Norway. henrik.lindhjem@nina.no

Abstract

With the current growth in broadband penetration, Internet is likely to be the data collection mode of choice for stated preference research in the not so distant future. However, little is known about how this survey mode may influence data quality and welfare estimates. In a first controlled field experiment to date as part of a national contingent valuation (CV) survey estimating willingness to pay (WTP) for biodiversity protection plans, we assign two groups sampled from the same panel of respondents either to an Internet or in-person (in-house) interview mode. Our design is better able than previous studies to isolate measurement effects from sample composition effects. We find little evidence of social desirability bias in the in-person interview setting or satisficing (shortcutting the response process) in the Internet survey. The share of "don't knows", zeros and protest responses to the WTP question with a payment card is very similar between modes. Equality of mean WTP between samples cannot be rejected. Considering equivalence, we can reject that mean WTP from the in-person sample is more than 30% higher. Results are quite encouraging for the use of Internet in CV as stated preferences do not seem to be significantly different or biased compared to in-person interviews.

Keywords: Internet; contingent valuation; interviews; survey mode; willingness to pay.

JEL Classification: Q51, H41

Introduction

One way the economics profession tries to support its self-proclaimed position as the only "hard" social science is by favouring new and sophisticated quantitative methods for recovering information from often poor data, over the less glamorous but essential groundwork of minimising and controlling survey errors in data collection. Economists valuing environmental goods using the contingent valuation (CV) method are generally no exception, though insights from psychology, survey methodology and other social sciences have penetrated the field to a larger extent than in other areas of economics – much due to the debate in the wake of the NOAA panel report on CV in natural resource damage assessments (Arrow et al. 1993). However, as the diminishing returns to yet another econometric method to analyse dichotomous choice data are setting in, it is worth pointing out – as do Boyle and Bergstrom (1999) – that potentially higher rewards may lie in gaining a better understanding of individual preferences in combination with improving CV data collection efforts to enable more robust insights from empirical analyses. Although current best practice CV studies do pay significant attention to questionnaire development and testing, the choice of data collection mode – mail, in-person, telephone, Internet[2] or a mix – is typically made with comparatively little evidence or consideration of its influence on how preferences are formed and stated. The issue becomes even more critical when considering that the CV literature has converged towards the view that preferences are discovered or constructed by the respondent during the data collection process (i.e. when the valuation questions are

[2] Computers have long been used in survey data collection both in combination with in-person interviews (so called CAPI – computer assisted personal interviewing) and telephone (CATI – computer assisted telephone interviewing). Our focus here is on self-administered surveys conducted on the Internet, usually while the respondent is in her home or workplace.

asked), rather than merely revealed or uncovered by it[3]. Traditionally, in-person interviews has been the recommended "gold standard" for CV (Mitchell and Carson 1989; Arrow et al. 1993). Mostly for reasons of lower cost, mail and to some extent telephone surveys are much more used in practice. The current trend in CV, like in other survey based research, however, is to collect data using the Internet. Sophisticated questionnaires can be delivered to large samples on record time at fairly low costs. Judging from the current growth in penetration rates, Internet has the potential to overcome the primary concern about population coverage and representativeness to become the mode of choice for survey data collection in the not so distant future (see e.g. Couper (2005))[4].

Several Internet-based CV studies of environmental goods, even ones such as Banzhaf et al. (2006) that may be considered best practice along other dimensions, have already been published or are in the pipeline (see e.g. Berrens et al. 2004; Thurston 2006;

[3] This has been an uncontroversial point in psychology and survey methodology for a long time. Survey methodologists make the point that data is a product of the collection process, i.e. generated at the time of the interview or completion of the questionnaire, rather than just "there" to be collected (implying that "data collection" is a misleading term) (Groves et al. 2004). More recently, environmental economists are also coming around to the view that preferences are constructed or learnt at the time of elicitation, at least when the preference object is unfamiliar to the respondent and/or she has little previous experience with it (McFadden 1999; Bateman et al. 2008; Carlsson 2010). This "constructivist" viewpoint does not necessarily mean that there is no "true" value or no stable and coherent preferences to be measured, only that economists need to be more sensitive to the fact that "the construction process will be shaped by the interaction between the properties of the human information processing system and the properties of the decision task, leading to highly contingent decision behaviour" (Payne et al. 1999:245). The survey mode is hence important in this regard.

[4] Almost a quarter of OECD inhabitants had broadband access in 2009, up from only 3.8 percent in 2002 (OECD 2010a). In EU, an average of 60.4% of households had internet access in 2008 (OECD 2010b). In Norway the figure for first half of 2009 was 86% SSB (2010). Dillman and Bowker's (2001) statement that the coverage problem in doing web surveys "is likely to persist in all countries in the world for the foreseeable future" sounds already dated (much like similar concerns about telephone coverage 40-50 years ago).

Lindhjem and Navrud 2009). Before the mass exodus proper starts from traditional survey modes to the Internet in CV and other stated preference methods, we think it is worth pausing to consider how this new mode may influence stated preferences and derived welfare measures for environmental goods. How does an Internet sample compare to a high quality in-person interview sample of the sort typically used in best-practice CV studies? Are Internet preferences biased and more unreliable or are the two modes equivalent? Which mode differences can be expected? Can mode effects be controlled within an acceptable range as we move more of the data collection to the Internet? These are the questions we attempt to answer in this paper.

In a controlled experiment as part of a national CV survey estimating willingness to pay (WTP) for proposed biodiversity protection plans, we assign two groups sampled from the same pre-recruited panel of survey respondents either to an Internet or an in-person interview mode. We can thus better control the effects of sample composition and measurement errors due to mode differences than the few previous studies that have attempted mode comparisons[5]. Both groups receive identical questionnaires administered during the same period by a professional survey firm. Adapting theoretical predictions and empirical findings from a broad survey methodology literature to the CV context, we investigate empirical differences between modes in our dataset and discuss reasons why such differences may occur. We limit our attention to elements of the CV survey of direct relevance to either estimation of WTP or judgements of the validity or quality of the data. We use both traditional tests of no difference between

[5] The two main sources of potential differences in stated preference results between survey modes are related to *methods of sampling* (i.e. affecting coverage error and non-response bias) and *questionnaire delivery* (i.e. affecting measurement error). The most important measurement error occurs when the same respondent provides different answers to survey questions that are worded the same across survey modes. Our focus here is on the measurement error due to mode – often termed the "survey mode effect".

modes and considerations of equivalence, i.e. whether mean WTP from the two modes for all practical purposes can be considered the same. Equivalence testing has a long tradition in pharmaceutical research when comparing whether two drugs have equivalent properties[6], and has also been used in survey mode research (Stanton 1998; Epstein et al. 2001) and benefit transfer in environmental economics (Kristofersson and Navrud 2005).

To our knowledge this is the first controlled comparison between Internet and in-person interview modes in stated preference research drawing samples from the same population. A few other studies compare Internet with in-person interviews (CV) (Marta-Pedroso et al. 2007; Nielsen In press), with mail (choice experiment) (Olsen 2009), with telephone recruited computer assisted survey (CV) (Dickie et al. 2007) or with both phone and mail (Taylor et al. 2009)[7]. However, the studies to date have generally compared modes with little conceptual guidance about which differences may be expected and why and typically confound sample effects with measurement effects. The only exception is the study recently commissioned by the USEPA, who has also begun to take seriously the issue of potential mode effects of Internet in stated preference surveys (Taylor et al. 2009). The general finding of the Internet comparisons,

[6] The analogy of comparing a new, cheaper and more convenient drug with functionally equivalent properties to and old drug, is quite striking in our case of Internet vs. face-to-face survey modes: "Dissatisfaction with the traditional null hypothesis has also emerged in an area of research in which the aim is not to establish superiority of one treatment or method over another, but rather to establish equality between the two methods. This type of research involves the testing of treatment innovations to determine if a new method achieves an equally effective outcome as the standard method but perhaps at lower cost or greater convenience" (Roger et al. 1993:553).

[7] In addition, Covey et al. (2010), Canavari et al. (2005) and van der Heide et al. (2008) (all in-person interviews), Banzhaf et al. (2006) (mail) and Li et al. (2004) and Berrens et al. (2003) (phone), all contain brief Internet comparisons based on surveys where the primary purpose generally was not to compare modes.

and the few that have compared other modes than Internet in CV[8], is that the choice of mode do affect value estimates and other parts of stated preferences, but that the reasons and direction are unclear[9]. We start in the next section by reviewing the theory and evidence of mode effects in survey research and CV. Based on this review part three derives our testable hypotheses. Part four gives a brief description of the survey design and data generation process. We find, as presented and discussed in part five that mean WTP from the in-person sample is not statistically different from the Internet sample. Finally, even though many survey mode effects are documented in the literature we are unable to discern clear indications in our data.

Survey mode effects and CV

Sources of survey mode effects

In their landmark book on CV Mitchell and Carson (1989) argued that the mode of choice for CV surveys is in-person interviews conducted in the respondent's home. Three main reasons were put forward for this: (1) the need to explain complex scenarios benefiting from use of visual aids with control over pace and sequence; (2) to motivate the respondent to exert a greater-than-usual effort to answer the WTP question; and (3) the importance of avoiding unit non-response for extrapolation to the population. They do, however, also acknowledge that telephone and mail may be suitable for surveying respondents who have familiarity with the good (e.g. recreational users). The NOAA panel concurred with this view and stated that it "believes it unlikely that reliable

[8] Notably Maguire (2009), MacDonald et al. (2010), Davis (2004), Ethier et al. (2000) and Legget et al (2003).

[9] A number of meta-analyses of the environmental valuation literature also document systematic, though not consistent, differences in welfare estimates depending on survey modes. One study found for example that high-response mail surveys gave lower WTP than low-response surveys (likely due to higher inclusion of less interested respondents) and both lower WTP than in-person interviews (Lindhjem (2007)).

estimates of values could be elicited with mail surveys. Face-to-face interviews are usually preferable, although telephone interviews have some advantages in terms of cost and centralized supervision" (Arrow et al. 1993:4608)[10]. The NOAA panel, however, recommends controlling for interviewer effects, especially social desirability bias, i.e. the tendency of respondents to edit their responses to appear in a more favourable light (DeMaio 1984). Schuman (1996) (the survey expert on the NOAA panel) defends and explains the NOAA recommendation of in-person interviews. Mail survey proponents, such as Don Dillman, strongly disagreed (see letter annexed in Schulze et al. (1996)). Schulze et al. (1996) called for more research comparing effects of different modes before definite recommendations for CV can be made. So, leaving effects of different coverage error and non-response bias between modes aside[11], what do we know about mode effects since the early 1990s?

Modes are likely to lead to different responses if they have different effects on the ways in which respondents come up with an answer. The response quality is determined by how carefully the respondent executes the process of understanding the question, retrieving information (including feelings, beliefs and knowledge about the environmental good), integrating information to form an overall judgement and formulating a response (Tourangeau et al. 2000). Two main human factors seem to be at work producing different responses between modes: one of a normative or sociological

[10] Note that the NOAA panel made recommendations for natural resource damage assessments, which may arguably be stricter than required for CV research more generally.

[11] Coverage error refers to differences in the definition of the population of inference due to the mode of data collection. Non-response bias is relevant when the (unobservable or observable) characteristics of people who prefer one mode to the other are correlated with the constructs we want to measure in the survey (e.g. WTP). The case where factors affecting the probability of response are correlated with the factors affecting the parameter(s) of interest is sometimes called sample selection bias (see e.g. Messonier et al. (2000) and Hudson et al. (2004) for a discussion of sample selection effects in CV).

nature and one of a cognitive or psychological nature (Dillman 2000). The former is related to how cultural norms are invoked differently across modes leading to culturally constrained responses. The main difference is between a self-administered situation and the involvement of an interviewer. The most important and well-documented mode effect in this regard, is according to Groves et al. (2004), social desirability bias. The extent of such responses seems to be closely related to two main factors: the degree of anonymity or "social distance", and trust, rapport or intimacy felt by the respondent. Social distance is minimised in an in-person interview conducted in the respondent's home, making socially desirable responses more pronounced. On the other hand, a great deal of interpersonal trust can emerge between an interviewer and the respondent in an in-home face-to-face interview, causing the respondent to be more honest resulting in less socially desirable responding. The net effect may be an empirical question (Holbrook et al. 2003). Since a CV survey consists of many different types of questions, some may be more susceptible to bias than others. As it is generally regarded as socially desirable to be in favour of environmental policies and to be an active recreationist, positive attitudes may be over stated and user days over reported in telephone or in-person interviews. Such biases may have implications for general assessment of the desirability of a proposed policy and for judging the validity of the CV data. The actual WTP question can be influenced by social desirability bias since it may be considered a "civic virtue" (much like voting) contributing to a common good. The effect may importantly depend on the payment format (open ended, payment card – PC, dichotomous choice – DC). DC is likely to be more susceptible to yea saying, a well-documented problem in in-person or telephone interviews, than in Internet or mail modes. However, for DC social desirability may be difficult to distinguish from the general tendency of people to answer affirmatively regardless of the content of the question (so-called "acquiescence"). For open-ended WTP questions (with or without

PC) it is less clear how social desirability works, though answering higher WTP may be the most likely response. For both WTP formats it is unclear a priori how social desirability may influence incentive compatibility and strategic bias[12]. It can perhaps be assumed that such effects are relatively neutral across survey modes. The degree of stated zero WTP and level of protesting (given zero) can be expected to be lower if social desirability effects are at work. This is of direct importance to the estimation of WTP. Other CV questions such as the degree to which the respondent has understood the scenario and whether he thinks the policy proposal is realistic – important for validity judgements of the data – may also not go free of bias. Finally, most of the background information collected in CV surveys will be truthfully reported regardless of mode (i.e. sex, age etc.), though some are typically not (especially income and education). Based on expected mode effects discussed above, different measures of social desirability for the whole or parts of the survey (e.g. as an index) or single questions can be constructed and tested

The second factor causing mode differences, the psychological, is related to individuals' cognitive processing of information and questions, in particular how aural and/or visual stimulus produces different responses across modes. To execute the response process well, respondents need to exert some degree of effort and in CV generally more so than in other surveys. Failure to put in the necessary effort to optimally answer a survey question, i.e. shortcutting the response process, leads to a satisfactory answer instead, or "satisficing" as coined by Krosnick (1991). Satisficing in the face of complex, lengthy questionnaires can take a myriad forms. Commonly observed effects are answering "don't know" or refusing (or generally more incomplete answers or item non-response),

[12] Differences in WTP response formats along these dimensions are considered important by economists, but are generally downplayed by psychologists (e.g. Green and Tunstall (1999)).

selecting the first reasonable response alternative, agreeing with assertions ("acquiescence"), non-differentiation (sticking to the same response category for a sequence of questions), endorsing status quo, "mental coin flipping" (random answers, if "don't know" is not offered as an option), choice of mid-points in rating scales, extremeness etc. All modes are likely to influence both the cost and the benefit side of the respondent's optimisation problem slightly differently. One of the proclaimed advantages of in-person interviews is the motivational effect of the interviewer. Green and Tunstall (1999) argue that in addition to practice (which is ruled out in most "one-shot" CV surveys), attention – which is more easily ensured by a motivated interviewer than in self-administrated surveys – will also improve respondent performance. The other advantage is that an interviewer can make it easier for the respondent to understand the information provided before stating his WTP and other responses. These two factors reduce respondent benefits of satisficing in interviews compared to the Internet mode. Similar to the discussion for social desirability bias, different types of CV questions will be susceptible to satisficing in different ways, with the WTP question an obvious victim. In a payment card, satisficing can conceivably lead to a tendency of picking the mid-point in the range (or perhaps less strongly: a narrower WTP distribution), more "don't knows" or even more zeros

Since little is actually known about the Internet as a survey tool it is sometimes assumed to be similar to mail surveys along the normative and social sources of mode effects discussed above (Dillman and Smyth 2007).

Comparisons of Internet and in-person interview modes in the CV literature

There is limited empirical evidence on social desirability bias and satisficing related to survey modes in the stated preference literature to further guide our empirical expectations. Marta-Pedroso et al. (2007) sample visitors to a beach for interviews (conducted by the authors) and Internet respondents recruited via an e-mail list. They found around the same share of zero WTP and protests for the two modes for an environmental preservation program in Portugal. Further, the mean WTP was found to be (much) higher for the interview than for the Internet sample (despite the fact that the Internet sample had much higher average income). The higher mean WTP in the in-person mode may be an indication of social desirability bias, although there are many confounding factors, including very different sample frames and sample compositions and a low 5 percent response rate for the Internet survey. There is also no consideration of the satisficing issue in the study. In a CV study of WTP for life expectancy gains from reduced air pollution, Nielsen (In press) makes a comparison between Internet and (in-home) in-person interviews recruited from two different sample frames. She finds significantly more protesting in the Internet sample, while the share of true zeros is similar between the modes. This finding indicates that people may find it socially easier to protest in the absence of the interviewer. Mean and median WTP is, however, not found to be different. The sensitivity to scope is also similar between modes, and there are few other indications that the Internet data somehow has lower validity or is more subject to satisficing strategies of respondents. The downside of the study is, in addition to the difference in sample frames, that the two surveys were carried out with more than one year lag.

In a choice experiment setting Olsen (2009) investigated preferences for protecting recreational use values from motorway encroachment in Denmark comparing a pre-recruited Internet panel sample with a general mail sample. Interestingly, he finds that the mail sample contains twice as many protestors as the Internet sample, though he concedes that this may just as well be due to self-selection into the Internet sample than real response differences. Comparing mean WTP Olsen (2009) concludes that it cannot be rejected that preferences from the two modes are identical. He then draws the, in our view, somewhat premature conclusion that "the fear of a potential survey mode effect is unfounded…". In a CV survey of reduced skin cancer risk Dickie et al. (2007) compare a sample recruited through a random digit dialling (RDD) procedure answering the survey on a computer in a central location with a sample of Internet panellists, collected three years later, answering on-line. Their results suggest lower quality of responses for the Internet survey, indicating greater satisficing (though the authors do not use this term). Internet respondents had more item non-response, rushed through the survey more quickly, indicated less awareness of the issue, took (perhaps) short cuts evaluating health risks, and failed a scope test of higher WTP for larger risk reduction. Higher motivation among the RDD respondents accepting to travel to a University campus for little compensation to complete the survey may be the most likely reason for this result. Dickie et al.'s (2007) design is unable to control many confounding factors, not least the large time lag of three years between the two surveys, so their conclusions are therefore speculative (which they also concede). Mainly addressing the issue of representativeness of two types of Internet samples[13] compared with a RDD telephone sample for political research, Berrens et al. (2003) also assess questions of environmental attitudes and WTP. They find that Internet respondents report more extreme attitudes and slightly lower share of yes votes for paying for a climate policy than phone sample respondents, a potential indication of social desirability bias. Importantly, Berrens et al. (2003) conclude that the analyst would make the same policy inference for the validity check of the data (e.g. that proportion of yes-votes decrease

[13] One sample comes from Harris Interactive (using an assembled panel of willing respondents to be sampled) and one sample from Knowledge Networks (using RDD-recruited households to a panel of Web-TV enabled respondents). These are the same sample types (and firms) also used by Berrens et al. (2004) and Li et al. (2005).

with bid price). Finally, Taylor et al. (2009), the study commissioned by USEPA, conduct the most thorough comparison to date of mode and sample effects of Internet, mail and phone in CV. They study WTP for air pollution reductions in the US and find that using either a (panel-based) Internet or mail survey produces more conservative values than the phone survey. They conclude that this result is due to social desirability bias, and further speculates that "...the apparent upward bias on the WTP due to the effects of social desirability in a phone survey would also be expected in a face-to-face survey." (Taylor et al. 2009: 5). However, this may not be obvious as, contrary to common belief, and those held by the NOAA panel, social desirability bias is often found to be larger in telephone than in in-person interviews, at least for sensitive questions (see e.g. Groves et al. (2004) or Jäckle et al. (2006)). In validity checks of the data, Taylor et al. (2009) do not find clear indications of lower quality Internet data due to satisficing or other effects, though the variance of WTP left unexplained was somewhat higher for the Internet sample.

In summary, there is fairly limited evidence of social desirability bias specific to in-person CV surveys that have been clearly distinguished from sample effects. Generally, the potential damping effect on WTP of interpersonal trust in an interview situation has been overlooked in the CV literature (as have any potential differences between on-site and in-home interviews). Even less has been said about satisficing effects. Both social desirability and satisficing are of course in many situations difficult to distinguish from each other, and from other potential psychological and sociological factors. A more thorough review of the use of Internet and other survey modes in stated preference research is provided by Lindhjem and Navrud (Forthcoming). In the next section we propose a few, simple indicators of mode effects of particular importance to CV surveys that will be tested in our data.

Hypotheses

Instead of investigating the whole CV survey instrument as if all questions are equally important, we believe it more fruitful to focus on satisficing and social desirability effects in the measurement of central variables for estimation of mean WTP and for the judgement of the validity of the data.

Satisficing & social desirability effects

We investigate two indicators of satisficing and social desirability bias adjusted from the survey literature to the first WTP question received by our respondents:

Hypothesis 1 (satisficing): The share of "Don't know" responses to the WTP question is *higher* for the Internet sample than for the in-person interview sample.

Hypothesis 2 (satisficing): The distribution of payment card responses has *lower* variance for the Internet than for the in-person interview sample[14].

Hypothesis 3 (social desirability): The share of stated zero WTP is *higher* in the Internet sample than in the in-person interview sample.

Hypothesis 4 (social desirability): The share of zero respondents that state reasons of protest is *higher* in the Internet sample than in the in-person interview sample.

The interpretation of Hypotheses 3 and 4 is that it may not only be less costly for the respondent to indicate zero WTP in the Internet survey, as some would see this as socially undesirable. But given that a respondent has stated zero, it may be an additional

[14] A stronger version of this hypothesis, increased tendency to choose midpoints in rating scales was hypothesised by Chang and Krosnick (2009).

hurdle to state a reason of (strong) protest in an interview situation, compared to a "safer" reason for stating zero. However, as has been discussed, the effects of the social desirability channels indicated in Hypotheses 3 and 4 may be dampened by the potentially induced response honesty resulting from interpersonal trust with an interviewer in the respondent's home.

Comparison of mean WTP

Of primary importance is the comparison of mean WTP between the two modes. Hypotheses 1-4 give indications of either social desirability effects or satisficing but the overall effect on WTP is undermined and an empirical question. A higher share of zeros in the Internet survey reduces mean WTP if the level of protesting is the same between samples (as such responses are typically taken out). However, we hypothesise that the share of protesting among zero respondents may also be higher in the Internet sample, so the share of true zeros could be the same in both samples – leaving a neutral mode effect. The effect on mean WTP of a higher level of "don't know" responses in the Internet sample is also unclear (such responses are also removed in WTP estimation). This is because the location in the WTP distribution of the additional share of "satisficers" in the Internet sample over the interview sample is unknown. If satificing is highest among low WTP-respondents, which may be likely[15], removing them in the Internet sample will increase mean WTP compared with the interview sample. Finally, the effect of Hypothesis 2 may go either way for the WTP comparison.

[15] It has been shown that for a range of indicators respondents with low education level is likely to have a higher tendency to satisfice (Holbrook et al. 2003). As education often is correlated with income, and income with WTP, the satisficing effects investigated here are more likely to be observed among low-WTP respondents.

The key question is if the two modes produce results that for all practical purposes can be considered equivalent, i.e. within a relatively small, predetermined bound. This is the primary convergent validity issue of interest[16]. Non-rejection of a traditional null hypothesis is not the same as demonstrating that the null is true. As has long been recognised, the null will often be rejected if sample sizes are large, "resulting in statistically significant differences that are substantively trivial" (Roger et al. 1993:553). Human behaviour in survey mode contexts (as in other contexts) can be said to be more elastic than allowed by a traditional non-difference test (see also footnote 2). Hence, it is important to determine if behaviour (in our case stated WTP) is "equivalent", not just (trivially) different. For this reason, we complement a traditional test of difference, with a test of equivalence, as noted previously. The agreed-upon standard adopted in pharmaceutical research for equivalence of two population means is +/- 20 percent (Rogers et al. 1993). 20-40 percent has been suggested by Kristofersson and Navrud (2007) for benefit transfer applications. We will take 20 percent as a starting point, considering other levels for sensitivity. We formulate the following two hypotheses:

Hypothesis 5a (classic null of no difference): Mean WTP is *equal* between the Internet and in-person interview samples.

Hypothesis 5b (non-equivalence of WTP): Mean WTP for the Internet sample is either *higher* or *lower* than for the in-person interview sample by 20 percent or more.

As discussed by Rogers et al. (1993) testing these two hypotheses can lead to four outcomes. First, if 5b is rejected and 5a confirmed, the analyst would conclude that no practically important difference exists between modes. Second, if both hypotheses were

[16] Since we in our survey do not have actual payment options, it is not possible to judge criterion validity of the two modes.

rejected, the conclusion would be that the difference is significantly larger than 0, but still trivial. This is the case where “too much” statistical power will tend to always reject the null, even if the difference is of little practical importance. Third, in the event that 5a is rejected, while 5b is confirmed, WTP are seen to be different and un-equivalent. Finally, if neither of the two hypotheses are rejected, the analyst would say that the “effect was not reliable enough to conclude either a sizeable difference or a reliably small difference” (Rogers et al. 1993:562).

Theoretical (construct) validity:

In addition to estimating WTP, the main population parameter of interest, we compare validity of the data for the two samples in terms of how WTP is related to other variables in a manner predicted by theory or as found in empirical research. Even if two modes may produce different response distributions for different types of explanatory variables in a CV survey, it is arguably their relationship with WTP that is important not the individual response distributions *per se*. We primarily investigate one dimension of validity: construct validity, as formulated in Hypothesis 6 below:

Hypothesis 6 (conformity of data with expectations): The relationship between WTP and commonly included explanatory variables is similar between modes in regressions.

Secondarily, we investigate the degree of internal scope sensitivity to the sizes of the two conservation plans. It would be difficult to judge whether higher scope sensitivity in the interview sample means social desirability bias or more valid data (both are conceivable), but a comparison may still be interesting.

Survey design and administration

Survey design and questionnaire content

The experiment was designed to test mode effects as part of a large multi-mode CV survey of increased biodiversity conservation in Norway, where the bulk of the data was collected over the Internet. There are government plans to increase the network of forest reserves from the current 1.4 percent of productive forest area to the minimum recommended by biologists of 4.5 percent to stem the loss of biodiversity (most of which are non-use related such as insects, fungi, mosses and plants). The questionnaire was developed following similar forest protection surveys well tested and tried in the Nordic context (see Lindhjem (2007)) and adopted to an Internet context following advice e.g. given by Dillman and Bowker (2001) and Dillman (2000). The instrument went through extensive testing in focus groups and two pilots using both Internet and in-person interviews.

The questionnaire first included questions about general use of government money for various ends to put the environmental good into a wider perspective and reduce potential focusing effects, before asking about the respondent's experience and use of forests in terms of recreational activities, and attitudes towards the perceived biological and aesthetical state of forests. Information was then presented about number and types of species, and the interplay between forestry practices, protection and development of the ecosystem functions and biodiversity in forests. Six colour photos of (neutral, "non-charismatic") endangered species and forest habitats were shown as well as pie and bar charts of number and percentage of species in all types of Norwegian habitats, including forests. The rather complex information was broken up with questions to activate the respondent and encourage response. After this information, respondents were presented current forest protection policy (status quo) and future plans. The environmental commodity was specified as two forest protection plans of either an increase to 2.8

percent (doubling) or to 4.5 percent (level recommended by biologists), presented to the respondent by advance disclosure (Bateman et al. 2004). The text was supplemented with colour maps of current and future forest reserves, and a table giving information about the size of new reserves, location of reserves and the likely improvements in the living conditions for main groups of species. The biological information was provided by a team of leading biologists and checked by foresters to ensure a balanced and realistic presentation of the status quo and future plans.

After the introductory sections, the respondents were reminded of their budget and given two open household WTP questions with the aid of a payment card (PC) for an annual, indefinite earmarked tax increase, starting with the small plan. We will use the responses to the first WTP question as the basis for testing our main hypotheses. The PC contained 24 amounts (ranging from 0 to NOK 15000) arranged on a non-linear scale in a table, including "don't know" (at the end). PC was chosen as response format over dichotomous choice (DC) to preserve sample efficiency. According to Boyle's (2003) review of the two response formats, it is far from clear that DC represents the better approach (as has been traditionally assumed since the time of the NOAA panel). The rest of the CV survey followed standard procedure; probing into why people answered zero or positive, checking their understanding and perceived realism of the scenario and WTP questions. The final part collected socio-economic background information[17].

Survey administration in the two modes

[17] The survey instrument is available from the authors upon request.

A randomly recruited panel of 35000 willing respondents, maintained by the professional survey firm TNS Gallup was used for the survey[18]. To the extent possible in a field experiment like this[19], confounding effects not related to survey mode was sought controlled as best as possible – as described in the following (partly based on considerations in Holbrook et al. 2003). First, two groups of respondents were interviewed either by in-person interview in their home or by Internet, which is better than subjecting the same respondents to both modes. Second, both samples were drawn randomly from the same population, i.e. the panel of respondents. Members of the panel with residence in the capital Oslo were chosen as the sample frame to reduce in-person interview costs. Third, respondents were not able to choose their preferred mode, but for practical reasons there were some small differences in recruitment to the survey. The in-person sample was recruited first by a standard e-mail invitation typically used for all surveys of this type to TNS Gallup's panel. It said that the survey (topic of which was not disclosed) would be conducted by in-person interview and those willing to participate were asked to reply to the mail. A random sample of those who replied was then contacted by phone to set up an interview time in the respondent's home at the respondent's convenience[20]. The panel mostly answers surveys on the Internet (and to a lesser extent mail and phone), so the recruitment procedure was made similar to a

[18] TNS Gallup uses no form of self recruitment, which is a common form of Internet survey recruitment (see Couper (2000) and Alvarez et al. (2003) for overviews of Internet survey types). This approach seems to be different from large survey firms such as Harris Interactive (US) and YouGov (UK), which assemble panels through many channels including self-recruitment by website advertisements with links etc.

[19] Moving such experiments out of the lab (or a central survey location, as used by Jäckle et al. 2006) gains something in terms of realism, but inevitably loses some degree of control over influencing factors.

[20] It was stated that the preferred location was in the respondent's home, but respondents who indicated that it would be practically difficult was offered to do the interview in TNS Gallup's central location downtown Oslo. This was done to reduce potential self-selection bias. Around 5 percent of the sample chose to have the interview centrally.

typical Internet survey. The Internet sample was then recruited from the panel using the same e-mail except that a weblink was included so respondents could enter the survey directly[21]. Since the panel contains background information about all members, the Internet sample was stratified based on age, sex and education to be as similar as possible to the in-person sample. This is an advantage normally not available in mode comparisons. Fourth, respondents that for some reason could not be interviewed by the suggested mode were not then assigned to the other mode (which is sometimes the case in practical mixed-mode surveys). Fifth, the questionnaire was identical between modes. The Internet survey was a page-by-page (not scrollable) design to make it easy to follow. The in-person interviews were conducted by nine experienced interviewers of varying age and sex, who were not informed about the purpose of the survey experiment. Questions were read to the interviewee with the aid of a small hand-held pocket computer and answers noted down by the interviewer on the screen[22]. For the most important questions, including the payment card, reply options were given on display cards with the same appearance as on the Internet to avoid well-known response order effects, which depend on whether alternatives are read or heard[23]. Identical maps, colour photographs and graphs were displayed from an interview folder in the same order as in the Internet survey.

[21] Ideally, respondents should first have been recruited and then randomly assigned to one of the two modes. However, for sake of realism, we chose to follow the common procedure used by the survey firm (e.g. it would have been unusual and potentially bad for response if panellists were to receive a survey invitation without information about how the survey would be carried out). We find it unlikely that our survey recruitment procedure biased the samples substantially according to respondents' survey mode preferences (see next section).

[22] This was the general rule, but if the respondent asked to read part of the information, she was given the opportunity to do so.

[23] So-called "recency effect" when the respondent picks a response option at the end of a list that is read by an interviewer (since the last options are contained in short-term memory), and "primacy effect" when the respondent picks something at the beginning of a list (more common when options are read by the respondent).

As a probe of social desirability bias, we asked interviewers to openly assess after the interview to what degree they thought the situation made it difficult for the respondents to say no to support the proposed program. We also asked interviewers whether they thought respondents had understood the WTP questions. These questions were phrased in neutral terms inviting an honest response from interviewers not implying any criticism against their handling of the interview (or any reference to social desirability bias or satisficing). The Internet survey forced respondents to answer questions before they could move to the next screen, so there was no item-non response in either mode. The average duration of the interviews was around 45 minutes, while completion times for the Internet survey were somewhat shorter, at around 25-30 minutes. As an indicator of respondent effort, we also measured the time it took Internet respondents to read and answer three parts of the survey: the introductory section with information on ecosystems, forests and endangered species; the section on current conservation policies; and the proposed policies and WTP questions[24]. Sixth, the surveys were conducted during the same time period in October and November 2007 to ensure preference stability and consistency between modes. Finally, the same token incentive payment to reply were given to both samples avoiding any related selection bias (Harrison and Lau 2009), and all respondents were interviewed individually[25]. Overall, the experimental design ensures a fairly tight control of the effects of survey mode for a typical CV survey of some length and complexity, without compromising realism for either mode. In the next section we discuss the composition of the two samples, before reporting results of the hypotheses tests.

[24] Unfortunately, information about time was not available for the in-person interviews.

[25] For the Internet survey it is impossible to be sure that other household members have not taken part or influenced the respondent. However, TNS Gallup informs respondents that they alone are supposed to answer the survey, likely giving a higher degree of control than in standard mail surveys.

Results and discussion

Samples and response rates

The response rates for Internet and in-person surveys were 59.7 and 75.4 percent, respectively[26], which compares favourably with similar surveys (even from pre-recruited panel respondents)[27]. These are final stage response rates and are therefore not adjusted for the response rates in the initial recruitment to the panel. For the purposes of our mode comparison experiment it is the final stage response rates that are relevant.

The socio-economic characteristics of the two modes, both for gross and respondent samples, are given in Table 1. All information (except for average household income) is taken from the database maintained by TNS Gallup about the panel and updated in the same year as the survey. Demographic variables are compared statistically between the two modes for both types of samples. Between the gross samples there are no statistical differences between age (distribution or average) and sex, but there are some differences between income and education distributions at the 10 percent level. This is indicated by the chi-square and t-statistics in column four. However, as can be seen comparing individual income categories, both samples are still fairly close. For the respondent samples (i.e. those from the gross sample who responded to the survey)

[26] 668 respondents first accepted to be interviewed, from which a sample of 398 was drawn. From this sample, 98 had to cancel appointments for various reasons, giving a final sample of 300, a 75.4 percent final-stage response rate. The original number of e-mail invitations for in person interviews were not given by TNS Gallup, precluding calculation of the more appropriate multi-stage response rate. However, TNS Gallup reports general response rates from the panel as high as 70-80 percent, indicating that the multi-stage response rate is unlikely to have been much lower than 40-50 percent.

[27] Berrens et al. (2004), for example, reports a response rate as low as 5.5 percent (completed web surveys to invitations sent).

there are no statistical differences between the two modes, except for the income distribution (which has lower significance now than for the gross samples) (see column seven in Table 1). However, a t-test rejects that average household income is statistically different between the respondent samples.

Table 1 Comparison of socio-economic characteristics between samples (percentages)

Socio-economic variables	**Gross samples**			**Respondent samples**		
	In-person (n=398)	**Internet (n=645)**	**Test stat. between modes**	**In-person (n=300)**	**Internet (n=385)**	**Test stat. between modes**
Internet use many times/day	89.7	87.6	t = 1.02	90.0	88.0	t = 0.81
Gender						
Male	46.8	49.4	t = -0.81	50.0	50.9	t = -0.23
Female	53.2	50.6		50.0	49.1	
Age			$\chi^2=0.37$			$\chi^2=0.46$
15-29	27.6	26.1		26.0	24.4	
30-44	36.7	38.0		38.3	37.9	
45-59	26.4	26.8		26.7	28.8	
60+	9.3	9.2		9.0	8.8	
Mean (number of years)	39.2	39.0	t = 0.23	39.5	39.9	t = -0.42
Household income (annual)			$\chi^2=16.3^{**}$			$\chi^2=11.4^{*}$
< 200 000	10.7	7.3		10.1	7.6	
200 000 – 399 999	27.4	20.7		26.9	20.2	
400 000 – 599 999	18.8	20.5		19.1	19.4	
600 000 – 799 999	15.7	19.9		16.1	20.1	
800 000 – 999 999	14.2	13.7		15.8	14.7	
> 1 000 000	8.9	9.6		8.1	9.7	
Not given	4.3	8.2		4.0	8.4	
Mean (Norw. Kroner)[¤]	-	-	-	631 449	585 487	t = 0.96

Education			χ^2=8.4*			χ^2=5.8
Primary (10 years)	6.1	6.3		6.3	5.5	
Vocational	29.3	35.3		27.8	33.9	
Secondary	19.2	19.5		20.7	20.4	
University (≤4 years)	26.0	18.9		24.4	18.0	
University (> 4 years)	19.4	20.0		20.7	22.2	

Notes: *,**,*** significance at 0.1, 0.05 and 0.01 levels, respectively. ¤ = As reported in the survey and estimated using midpoints in indicated in much more detailed income categories than for the income information in the gross samples. Pearson's chi-square test used to compare frequency distributions.

Further, the rate of Internet use is not different between samples, i.e. there is no tendency that those who use Internet less have to a larger extent responded in the interview mode. Non-response among groups according to age, income and education seems very similar for the Internet an in-person modes, respectively. This means that the type of survey mode does not overall seem to have influenced whether people participated and responded or not – both gross and respondent samples show no large deviations that are likely to confound the measurement effects of mode. We therefore proceed by testing our hypotheses and investigate validity of the data without weighing the samples by socio-economic characteristics or conducting further investigation of non-response effects[28].

Satisficing & social desirability

We start by reporting the results from the satisficing hypotheses (H1 & H2). When asked the WTP question it is likely that satisficing would lead to a higher share of "don't know" responses indicated in the payment card (PC) for the Internet survey. However, the data rejects this hypothesis: 11 percent of Internet respondents and 8

[28] A more comprehensive analysis could have included both running a Heckman sample selection model (Heckman 1979) (e.g. as conducted by Banzhaf et al. 2006) or weighing samples by demographics.

percent of the interview respondents state “don’t know”, a difference in the expected direction, though not significant on the 10 percent level (see row three in Table 2, and Figure 1 below). The second, more explorative hypothesis that PC responses are clustered more closely together in the Internet survey, expressed as lower variance for the WTP distribution, is also rejected using a likelihood ratio test for the parametric WTP model explained in the next section (see footnote 31) (row four in Table 2).

Table 2 Test results for indicators of satisficing and social desirability

Hypotheses: Satisficing & Social desirability		**Sample modes**		**Mode comparison**	
		Interview **(n=300)**	**Internet** **(n=385)**	**Test statistic**	**Result (p<0.1)**
H1	Share of “don’t knows” *higher* on web	8.0%	11.1%	t = 1.38	Rejected
H2	WTP variance *lower* on web	σ = .978	σ = 1.26	χ^2 =14.27[a]	Rejected
H3	Share zero responses *higher* on web	19.3%	18.9%	t = -0.12	Rejected
H4	Share protest responses *higher* on web				
	All except can’t afford or no value	90.65%	88.06%	t = -0.64	Rejected
	Tax, gov’t or responsibility	74.77%	70.90%	t = -0.66	Rejected

Note: a: Likelihood-ratio test of equality of standard error, sigma (σ), as explained in 31.

Probing further into the issue of satisficing, we checked whether a higher share of Internet respondents found it “very hard” to answer the WTP question. 25 percent of Internet respondents and 17 percent of interview respondents stated this, but the difference is not significant. We also found significantly higher degree of “don’t knows” to the WTP question for respondents with the lowest education compared with higher education respondents within both modes, as expected from theory and previous studies. There is also a difference *between* modes (26.3 percent of low education interview respondents stated “don’t know” vs. 33.3 percent for the Internet sample).

Time spent reading information and answering questions in the Internet survey may say something about the effort people expend and the degree of satisficing. The median time spent on the introductory section about ecosystems, forests and endangered species was 90 seconds, while median times to complete the two sections on current policies and new policies including the WTP question were 105 seconds each. Running two simple probit models using either "don't know" or zero response to the WTP question as the dependent binary variable (results left out for sake of brevity), we find highly negative and significant coefficients for the time spent by respondents answering the WTP question[29]. This means that the less time respondents spend on the WTP question, the more likely they are answering don't know or zero. This is an indication that both these response types may result from satisficing strategies rather than a thorough consideration of the WTP question. We also include the time variables in the modelling of WTP below.

Moving to indicators of social desirability, we first test the hypothesis that the share of zero PC responses is higher in the Internet survey (H3). No such difference is found in the data: both shares are close together at 19.3 and 18.9 percent for the in-person and Internet modes, respectively (see row five in Table 2 and Figure 1 below). Hence, there is no evidence that the interview situation makes it socially harder for respondents to state a zero response, an important finding for CV research. Second, we tested whether two types of protesting which slightly different interpretation for social desirability were more common in the Internet survey (H2). When answering zero respondents would be

[29] For don't know responses the result is robust at the 10 percent level for times from 0-600 seconds (i.e. 10 minutes), which includes 95.6 percent of responses. For zero responses the result is robust at the 5 percent level for times from 0 to 4000 seconds (67 minutes), which includes 98.7 percent of responses. A few responses were excluded for which measured time was either very large (indicating that the respondent may have left the computer to resume at a later time) or negative (indicating some computer clock problem or faulty measurement).

asked in standard CV fashion to state up to two reasons from a list of possible reasons to enable identification of protest responses. A strict interpretation of protest would be to include all those who state zero WTP even if the good has a positive value to them and they are not prevented from paying by an income constraint. Using this interpretation, of all stated reasons for zero WTP only 9.35 and 11.94 percent for the interview and Internet samples, respectively, were true zero reasons. This leaves shares of protest that are not statistically significant (see row seven in Table 2). Speculating that social desirability effects may work differently for different types of protest reasons, we conducted a second classification of protest responses. Protest reasons that may carry a perceived higher "social punishment" in the interview situation, e.g. related to taxes ("too high") and responsibility for causing or solving the problem ("it's a government responsibility", "those who destroy habitats should pay"), were distinguished from idealistic reasons ("it is wrong too value biodiversity in monetary terms") or response difficulties ("too difficult to come up with a value"). The latter types of responses are perhaps easier to state with "a straight face". Classifying only the former types of responses as strict protest gave somewhat surprisingly a share of 74.77 percent protest in the interview sample and 70.90 in the Internet sample (row eight in Table 2). In summary, the assessment of both zero responses and protesting give no evidence in the data of social desirability bias.

We also conducted a more causal inspection of indications of socially desirable responding to four potentially susceptible non-WTP questions. The first two questions, related to whether or not the respondent had recreated in a forest the last year and if so, how many times last month, gave no difference in response distributions. Further, no discernable differences were observed in respondents' self-assessment of knowledge of biodiversity loss or their attitude towards doing something about the problem. These results run contrary to those of Legget et al. (2003) who find indications of socially desirable responding to questions of whether the respondent had visited the site before, if she thought the visit was too short, if she enjoyed it and whether the site was the primary purpose of the trip. Such evidence (or lack thereof) is more important for our judgement of whether socially desirable responding is prevalent in a survey – perhaps also spilling over to the WTP question – than for the use of the results from these questions *per se*.

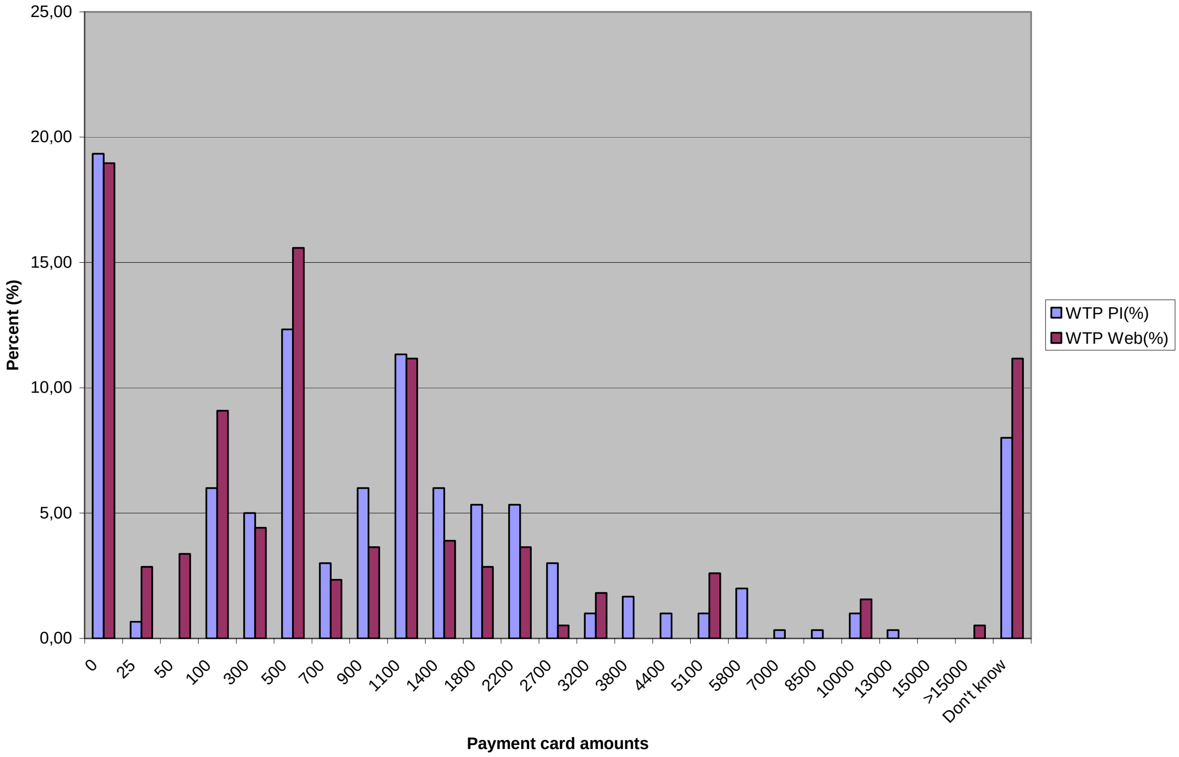
Finally, we included an additional probe of social desirability bias, asking interviewers to openly assess after the interview to what degree they thought the situation made it difficult for the respondents to say no to support the proposed program. 14.5 percent of interviewers answered "to some or to a large degree", while 67 percent answered "to small degree or not at all" – indicating a fairly limited degree of perceived pressure in the interview situation. We test this indicator in the WTP models in the next sections.

Overall then, little evidence has been found in our data for the hypotheses of social desirability bias and lower level of satisficing in the in-person interviews. The next step is to compare mean WTP between modes.

Comparison of mean WTP

The WTP distribution as indicated by respondents in the PC is depicted in Figure 1 for the two modes, including zero and "don't know" at opposite ends of the diagram. No obvious differences between modes can be discerned from Figure 1. To test Hypotheses 5a and 5b (H5a & H5b), of either difference or equivalence of mean WTP between modes, we start by estimating mean WTP following standard parametric procedures for interval PC data discussed in Cameron and Huppert (1989). Since the stated WTP amounts have a skewed distribution with the familiar long right tail, a log-transformation of WTP was applied[30].

Figure 1 Household WTP distribution as indicated in payment card. Norwegian Kroner, annual amounts for an indefinite period.



Since both levels of protest and zero responses have been shown not to be statistically different between modes and because determining true zeros is somewhat controversial, we exclude all zeros for simplicity along with "don't know" responses from our estimation and focus on positive WTP responses. This has no practical importance for our conclusion. Further, no WTP responses that could be considered extreme were

[30] Mean WTP from this model is given by E(WTP)=exp(a $+\sigma^2/2$), where a and σ are the estimated parameters from the lognormal model. True WTP lies between the lower limit – as indicated by respondents in the PC – and the upper limit of each PC interval.

identified and very little item-nonresponse (e.g. for income) in both modes ensure almost full samples. Mean WTP is given in Table 3.

Table 3 Comparison of mean WTP between modes. WTP in Norwegian Kroner.

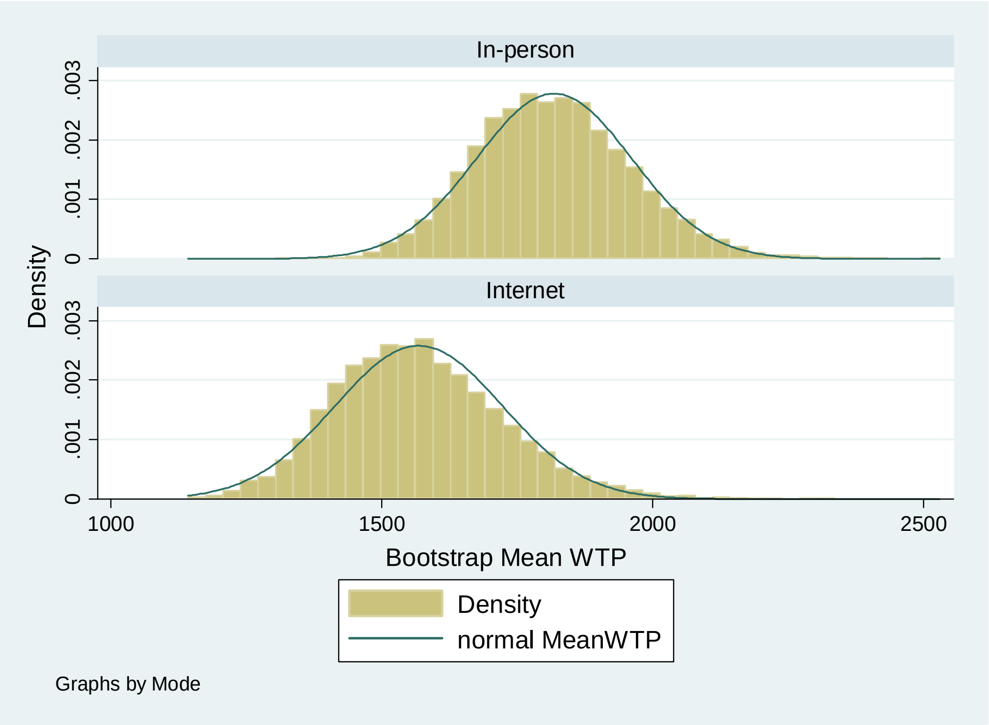
Hypothesis	**Mean WTP Interview: (95% CI) (n=218)**	**Mean WTP Internet: (95% CI) (n=269)**	**Mode comparison result (p<0.1)**
H5a Equality of mean WTP	1819 (1539, 2100)[a]	1566 (1261, 1871)[a]	Non-rejection

Notes: Estimated using interval regression in STATA 9.2. a: 95% confidence intervals calculated using 10000 bootstrap draws with replacement, following Efron (1997). 1 Norwegian Krone (NOK) = ca 0.125 Euro at time of study.

The mean for the interview sample is somewhat higher at NOK 1819 than the NOK 1566 for the Internet sample[31]. We calculate 95 percent confidence intervals around the respective means based on a bootstrap (10000 draws with replacement) from each of the sample distributions. Since the confidence intervals are overlapping we cannot reject hypothesis 5a that mean WTP are equal between modes on the 5 percent level. The bootstrap distributions of means are depicted for both samples in Figure 2, showing the somewhat higher mean for the in-person interview sample.

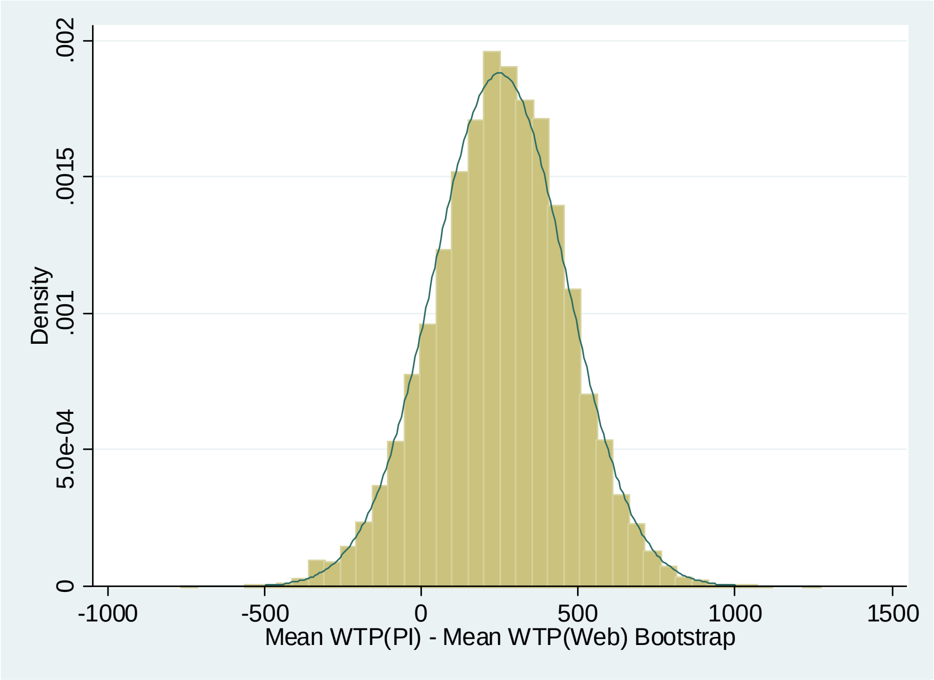
[31] For comparison, the conservative non-parametric mean (median) WTP based on using the WTP amounts indicated in the PC, i.e. as shown in Figure 1 (rather than the mid-points in each interval) are NOK 1599 (1100) and NOK 1361 (500) for the interview and Internet modes, respectively.

Figure 2 Distribution of bootstrapped mean WTP from the two samples (10000 draws with replacement)



However, as we have argued, failing to reject the traditional null can in our case not be constructively interpreted as confirming convergent validity of WTP estimates between modes (in the same way a rejection cannot meaningfully be taken as evidence against convergent validity). Instead, we investigate whether the difference between means is of practical importance, i.e. larger than a predetermined bound (Hypothesis 5b - H5b). To test this hypothesis we combine the two bootstrapped mean WTP distributions in Figure 2 into a single distribution of the differences in mean WTP for the two modes (see Figure 3). First, we can observe that most of the distribution is larger than zero. However, since only 87.95 percent is, we cannot reject H5a at the 10 percent level. This is also shown in that the 95 percent confidence intervals around the means in Table 3 are overlapping.

Figure 3 Distribution of bootstrapped mean WTP(Interview) – mean WTP (Internet)



We conduct the same simple non-parametric procedure to test how much of the distribution is outside different equivalence intervals. Table 4 displays results. First, testing whether mean WTP for the Internet sample is higher or lower than 20 percent of mean WTP for the interview sample (i.e. ± NOK 364) leads to non-rejection since around 30 percent of the distribution is contained outside this bound (see row three in Table 4). In other words, observing a sample difference between means of NOK 253, we cannot reject that the population difference may be larger than 20 percent. However, if we a priori deem a difference of 30 percent between means as acceptable for equivalence, the hypothesis of non-equivalence can be rejected at the 7.55 percent level (row six in Table 4). The cut-off point between rejection and non-rejection is 28 percent difference, at the 10 percent confidence level (row six).

Table 4 Test of non-equivalence of mean WTP between modes

Hypothesis:		Equivalence criterion (EC): WTP difference (NOK)	Percent of WTP diff. distribution outside EC	Mode comparison result (p<0.1)
H5b	Non-equivalence, 10%	± 182	66.26	Non-rejection
	Non-equivalence, 20%	**± 364**	**30.17**	**Non-rejection**
	Non-equivalence, 25%	± 455	16.04	Non-rejection
	Non-equivalence, 28%[a]	± 511	9.99	Rejection
	Non-equivalence, 30%	± 546	7.55	Rejection
	Non-equivalence, 40%	± 728	1.10	Rejection

Notes: a: 28% is the difference between means, which allows rejection at the exact 10 percent level.

If we keep to the 20 percent equivalence level, we are unable to reject any of our hypotheses 5a or 5b. This means we cannot conclude "either a sizeable difference or a reliably small difference" (Rogers et al. 1993: 563) between modes. However, the sensitivity analysis shows that increasing the acceptable level of difference to 30 percent would comfortable reject H5b. Hence, the equivalence test adds useful information to the conclusion given from the standard hypothesis test of no difference.

Theoretical validity

Our final hypothesis is whether the relationship between WTP and common explanatory variables is similar between modes, i.e. a type of theoretical or construct validity check. Table 5 presents results of four double log interval regression models. Model 1 and 3 include the same socio-economic, use, attitude and other variables for both modes for sake of comparison. Models 2 and 4 add to these mode specific variables, to be explained below[32].

[32] Based on the results of a likelihood ratio test, we do not run pooled models. The likelihood ratio statistic is q=-2[$\log L_{PooledAB}$ – ($\log L_A$+$\text{Log}L_B$)~$\chi 2$ (d.f.), where $\log L_A$ and $\log L_B$ refer to the log likelihood values from the estimated models for WTP for individual samples (without covariates), and $\log L_{PooledAB}$ is the likelihood value for a pooled model. Running the pooled model without a sample dummy yields a test static of 32.99, which allows us

Table 5 Estimation results for in-person interview and Interview modes.

Independent variables		**Interview sample**		**Internet sample**	
		Model 1	**Model 2**	**Model 3**	**Model 4**
Socio-economic:					
Sex[a]	1 if male	.157 (.133)	.112 (.137)	.194 (.143)	.317** (.144)
LnAge[a]	>15 years of respondent	.301 (.199)	.302 (.214)	.464** (.213)	.458** (.211)
LnInc	Hhld income, mid-points	.163* (.092)	.160* (.092)	.214** (.107)	.216** (.107)
Eduhigh[a]	1 if > 4 years univ. educ.	-.138 (.156)	-.155 (.161)	-.057 (.173)	-.012 (.170)
Edulow[a]	1 if only primary educ.	-.138 (.156)	.230 (.332)	.145 (.348)	.351 (.343)
LnHhld[a]	# adults & children	-.227 (.194)	-.222 (.202)	-.274 (.233)	-.143 (.234)
Use, attitudes, other:					
Member	1 if memb.of nature org.	.681*** (.196)	.686*** (.197)	.937*** (.304)	.825*** (.297)
Use	1 if forest visit 12 mths	.266 (.322)	.338 (.344)	.253 (.301)	.303 (.309)
LnTrips	>15 forest, 1 mth	.001 (.085)	-.004 (.086)	.102 (.092)	.085 (.091)
Nouse	1 if not to use reserves	-.393** (.164)	-.410** (.171)	-1.048*** (.316)	-1.143*** (.3181)
Attax[a]	1 if agree w. taxes	.218 (.139)	.250 (.143)	.157 (.175)	.177 (.172)
Difficult	1 if hard to answer WTP	-.066 (.172)	-.087 (.182)	-.402** (.190)	-.364** (.187)
Mode specific:					
LnTime1[b]	Seconds read.intro. info				.093 (.084)
LnTime2[b]	Sec. reading policy info				-.113 (.139)
LnTime3[b]	Seconds answering WTP				.428*** (.130)
IntUnd	Understand WTP quest.		.050 (.138)		
IntPress	Hard to say “no” interv.		.106 (.303)		
Int1	Interviewer #1		-.170 (.924)		
Int2	Interviewer #2		-.308 (.910)		
Int3	Interviewer #3		-.216 (.944)		
Int4	Interviewer #4		.056 (.966)		
Int5	Interviewer #5		-.093 (.912)		
Int6	Interviewer #6		-.229 (1.022)		
Int7	Interviewer #7		-.389 (.349)		
Int8	Interviewer #8		-.180 (.346)		
IntAge	Interviewer age		.000 (.025)		
IntSex	Interviewer gender		-.008 (.319)		
Constant		3.705*** (1.154)	3.747** (1.918)	1.990* (1.237)	-.260 (1.347)
Log Likelihood		- 534.04	-531.22	- 701.04	-673.50
N[c]		206	206	268	260

Notes: *,**,*** significance at 0.1, 0.05 and 0.01 levels, respectively. Dependent variable is WTP intervals from the payment card. Ln: means log transformations. a. Variable information taken from respondent panel database updated in 2007. Other variables are from the CV survey. b. Time use information only available from Internet

to reject that both parameters are equal at the 1 percent level. Running the same model with a sample dummy yields 14.27, which means we can also reject that the standard errors are the same at 1 percent level – meaning that the two samples cannot be pooled survey. c. A few respondents did not state income, so these observations have been excluded. Interval regression in STATA Version 9.2. used.

The first point to note is that there are no great differences between Models 1 and 3 in terms of signs of coefficients or degree of significance. The coefficients on income and membership in a nature conservation organisation are positive and significant for both modes, as expected. Further, if the respondent has no intention to use any new forest reserves ("Nouse"), he tends to provide a lower WTP, also as expected. The coefficients on current use of forests (typically not reserves) for recreation as a dummy ("Use") or number of trips ("LnTrips") are small and insignificant. This is not necessarily surprising as very few people actually use existing forest reserves (as they are remote and inaccessible), so may realise most of the value will be related to non-use. Older respondents state higher WTP (significantly so only for the Internet sample), while gender and education levels have no clear effect on WTP. On the basis of the simple comparison of the two models, we cannot reject that the degree of construct validity is similar between the two modes using a selection of commonly included explanatory variables – and no different from regression results typically observed in the CV literature (e.g. in Banzhaf et al. 2006).

To complement the analysis of social desirability and satisficing above, we included some additional variables. First, whether respondents indicated that they thought it was "very hard" to answer the WTP question is included as a dummy variable, "Difficult". As noted earlier, more respondents in the Internet sample held this view. Interestingly, respondent difficulty seems to translate into significantly lower WTP only in the Internet mode. This result should be interpreted with caution, but it does indicate that if WTP questions or scenarios can be made easier to follow also for self-administered surveys, WTP differences between modes may narrow. Further, we included dummies for interviewers and their age and gender, to control for potential interviewer effects, such as those found by Bateman and Mawby (2004) or Loureiro and Lotade (2005). None of these coefficients are significant, indicating fairly consistent interviewing and no specific bias across the 9 interviewers that did the bulk of the interviews. Finally, as noted, we measured the time it took Internet respondents to complete three separate sections, included as variables, “Time1”, “Time2 and “Time3”. Interestingly, the first two dummies are not significant, but the third is. The more time Internet respondents spent thinking about the WTP question the higher is the WTP they state. This could trivially be because interested respondents spend more time on surveys and state higher WTP, but it is not clear since time spent on the other parts of the survey has no effect on WTP. Holgraves (2004) found that socially desirable responding was related to longer response times (not directly related to CV and WTP). However, this sounds unlikely to be the case for our Internet mode. Unfortunately, we have no comparable time measurements for the in-person interviews.

Finally, we checked whether people increased their WTP when the alternative and larger protection plan was offered. The shares of respondents going up, staying at the same level or reducing their bid are roughly equal across the two modes. The shares for the interview sample are 47.4, 51.6 and 0.9 percent and for the Internet sample 47.5, 48.3 and 4.2 percent, respectively. Internal scope validity is therefore similar between modes, although it may not be a very precise indicator of reliable CV data (Amiran and Hagen 2010) (or social desirability).

Concluding remarks

In a controlled CV field experiment we have conducted the first test of whether responses and stated preferences are different between collecting data using the Internet or in-person interviews in the respondents’ home. Since both samples are drawn from the same panel of willing respondents, we are better able than previous studies to isolate effects of the survey mode from sample composition effects. Checking in particular for indications of social desirability bias and shortcutting of the response process (satisficing), both well-documented effects in the broader survey literature, we find little evidence in our data. We find that the extent of "don't know", zeros and protest responses to the WTP question (with a payment card) is similar between modes. There is also no tendency of payment card responses being more closely clustered together in the Internet mode. Mean WTP is somewhat higher in the interview sample, though we cannot reject that mean WTP in the two modes are equal on the 10 percent level. We also consider equivalence, i.e. whether it can be rejected that the WTP difference is larger than a practically, trivial predetermined bound. We can reject that the difference is more than 30 percent, but fail to reject an equivalency bound of 20 percent on the 10 percent level. For practical purposes it is useful also to conduct the equivalency test, as failure to reject the traditional null hypothesis of no difference cannot uncritically be taken as evidence of convergent validity between modes. Kristofersson and Navrud (2007) argue in benefit transfer applications that the level of required accuracy should depend on types of policy uses (e.g. lower accuracy is acceptable for cost-benefit analysis than for natural resource damage assessments). They suggest that differences of 20-40 percent may be acceptable, depending on the context. Equivalency testing becomes even more topical when considering that the use of Internet in experimental economics is likely to grow, enabling large, low cost split-samples, and tests that will typically find significant, though often practically trivial, differences between treatments. Finally, we check whether WTP vary in similar ways with common explanatory variables for both modes. The two modes show the same degree of construct validity for different WTP model regressions. Further, we find no evidence that interviewers influence WTP differently (i.e. no interviewer effects).

We have considered mode effects in our data documented in a fairly broad literature in survey methodology, psychology and sociology. We are keenly aware of Jason Shogren's general warning ...to experimental economists that: "economists venturing into this cognitive minefield alone will end up fifty years behind the psychologist's times" (Shogren 2005). Hence, although we have focussed on social desirability and satisficing – and find little evidence of such effects of direct relevance to estimation of WTP – we acknowledge that there are many cognitive processes and decision heuristics at work we cannot control for in a field setting. Further, we are cautious of generalisation, as our CV survey relates specifically to a complex, environmental good of potentially high non-use values in a European country. Results may not directly extend to choice experiment settings, goods with higher use values, or countries with very different cultures. Social desirability bias is for example likely to be more pronounced in cultures where it is not considered "polite" to disagree etc. (see e.g. Karp and Brockington (2005) for a voting example and Ehmke et al. (2008) for an international comparison of hypothetical bias where cultural differences matter).

Given the complexity of our survey and good and the lack of clear, documented social desirability bias or interviewer effects, in-person interviews is likely to be the preferred mode – as also noted by the NOAA panel. "One shot" in-person interviews is also a compromise between mail, phone and Internet and the more deliberative approaches recently introduced in CV to facilitate a better learning or construction of preferences for complex and unfamiliar goods (see e.g. MacMillan et al. (2006), Bateman et al. (2008) or Lienhoop and MacMillan (2007)). However, for reasons of cost, convenience and opportunities for better designs, Internet, either as stand-alone applications or as the primary mode in mixed-mode stated preference surveys, is set to grow tremendously. Whereas the coverage and representativeness concerns about Internet are likely gradually to be reduced in Western countries (much like concerns over phone coverage some decades ago), potential measurement differences between modes will remain. In this respect, our results are quite encouraging in that values derived using the Internet seem not to be significantly different or less reliable compared to in-person interviews. Further, if anything, our results show that the Internet mode gave slightly lower WTP. Since we do not know the true WTP of the respondents, it is important to estimate those values conservatively. Finally, this is a humble, first attempt to compare Internet with in-person interviews. More research is necessary not only to document mode effects, but also to better pin down and understand their causes, so potential measurement biases can be controlled within acceptable ranges in future CV applications.

Acknowledgements

This paper is based on surveys conducted for the project: "Non -timber forests goods and local economic development: An untapped potential?" (Project no. 173230), financed by the Research Council of Norway. We have benefitted from comments from Olvar Bergland, Bob Berrens, Kevin Boyle, Mark Dickie, Shelby Gerking, Karine Nyborg, Randall Rosenberger, V. Kerry Smith and participants at the W-2133 -meeting "Benefits and Costs of Natural Resource Policies Affecting Public and Private Lands”; The Western Association of Agricultural Experiment Station Directors (WAAESD) February 24-26 2010, Tucson, USA and World Congress of Environmental and Resource Economists, Montreal, Canada June 28 - July 2 2010. The usual disclaimer applies.

References

Alvarez R M, Sherman R P and VanBeselaere C (2003) Subject acquisition for Web-based surveys. Polit. Anal. 11(1): 23-43

Amiran E Y and Hagen D A (2010) The Scope Trials: Variation in Sensitivity to Scope and WTP with Directionally Bounded Utility Functions. Journal of Environmental Economics and Management 59(3): 293-301

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

Banzhaf H S, Burtraw D, Evans D and Krupnick A (2006) Valuation of natural resource improvements in the Adirondacks. Land Economics 82(3): 445-464

Bateman I, Cole M, Cooper P, Georgiou S, Hadley D and Poe G L (2004) On visible choice sets and scope sensitivity. Journal of Environmental Economics and Management 47: 71-93

Bateman I J, Burgess D, Hutchinson G H and Matthews D I (2008) Learning design contingent valuation (LDCV): NOAA guidelines, preference learning and coherent arbitrariness. Journal of Environmental Economics and Management 55: 127-141

Bateman I J and Mawby J (2004) First impressions count: interviewer appearance and information effects in stated preference studies. Ecological Economics 49(1): 47-55

Berrens R P, Bohara A K, Jenkins-Smith H, Silva C and Weimer D L (2003) The advent of Internet surveys for political research: A comparison of telephone and Internet samples. Polit. Anal. 11(1): 1-22

Berrens R P, Bohara A K, Jenkins-Smith H C, Silva C L and Weimer D L (2004) Information and effort in contingent valuation surveys: application to global climate change using national internet samples. Journal of Environmental Economics and Management 47(2): 331-363

Boyle K J (2003) Contingent valuation in practice. In: P. A. Champ, K. J. Boyle and T. C. Brown (ed), A primer on nonmarket valuation. Kluwer Academic Publishers,

Boyle K J and Bergstrom J C (1999) Doubt, doubt, and doubters: The genesis of a new research agenda? In: I. Bateman and K. G. Willis (ed), Valuing Environmental Preferences. Oxford University Press,

Cameron T A and Huppert D D (1989) Ols Versus Ml Estimation Of Non-Market Resource Values With Payment Card Interval Data. Journal of Environmental Economics and Management 17(3): 230-246

Canavari M, Nocella G and Scarpa R (2005) Stated Willingness-to-Pay for Organic Fruit and Pesticide Ban: An Evaluation Using Both Web-Based and Face-to-Face Interviewing. Journal of Food Products Marketing 11(3): 107-134

Carlsson F (2010) Design of Stated Preference Surveys: Is There More to Learn from Behavioral Economics? Environmental and Resource Economics 46(2): 167-177

Chang L and Krosnick J A (2009) National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality. Public Opinion Quarterly 73(4): 641-678

Couper M P (2000) Web surveys - A review of issues and approaches. Public Opin. Q. 64(4): 464-494

Couper M P (2005) Technology trends in survey data collection. Soc. Sci. Comput. Rev. 23(4): 486-501

Covey J, Robinson A, Jones-Lee M and Loomes G (2010) Responsibility, scale and the valuation of rail safety. Journal of Risk and Uncertainty 40: 85-108

Davis J (2004) Assessing Community Preferences for Development Projects: Are Willingness-to-Pay Studies Robust to Mode Effects? World Development 32(4): 655-672

DeMaio T J (1984) Social desirability and survey measurement: A review. In: C. F. Turner and E. Martin (ed), Surveying subjective phenomena: Volume 2. New York: Russel Sage,

Dickie M, Gerking S and Goffe W L (2007). *Valuation of Non-Market Goods Using Computer-Assisted Surveys: A Comparison of Data Quality from Internet and RDD Samples*

Dillman D (2000) Mail and internet surveys: the tailored design method. John Wiley & Sons, Inc,

Dillman D and Bowker J M (2001) The WEB questionnaire challenge to survey methodologists. In: U.-D. Reips and M. Bosnjak (ed), Dimensions of Internet Science. Lengerich, Germany: Pabst Science Publishers, pp. 159-178

Dillman D A and Smyth J D (2007) Design Effects in the Transition to Web-based Surveys American Journal of Preventive Medicine 32: S90-S96

Ehmke M D, Lusk J L and List J A (2008) Is Hypothetical Bias a Universal Phenomenon? A Multinational Investigation. Land Economics 84(3): 489-500

Epstein J, Klinkenberg W D, Wiley D and McKinley L (2001) Insuring sample equivalence across internet and paper-and-pencil assessments. Computers in Human Behavior 17: 339-346

Ethier R G, Poe G L, Schulze W D and Clark J (2000) A comparison of hypothetical phone and mail contingent valuation responses for green-pricing electricity programs. Land Econ. 76(1): 54-67

Green C and Tunstall S (1999) A psychological perspective. In: I. Bateman and K. G. Willis (ed), Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries. Oxford University Press,

Groves R M, Fowler jr F J, Couper M P, Lepkowski J M, Singer E and Tourangeau R (2004) Survey Methodology. Wiley,

Harrison G W and Lau M I (2009) Risk attitudes, randomization to treatment, and self-selection into experiments. Journal of Economic Behavior & Organization 70: 498-507

Heckman J J (1979) Sample selection bias as a specification model. Econometrica 47(1): 153-161

Holbrook A L, Green M C and Krosnick J (2003) Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparison of respondent satisficing and social desirability response bias. Public Opinion Quarterly 67: 79-125

Hudson D, Seah L, Hite D and Haab T C (2004) Telephone presurveys, self-selection, and non-response bias to mail and internet surveys in economic research. Applied Economics Letters 11(237-240)

Jäckle A, Roberts C and Lynn P (2006). *Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes*. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project,

Karp J A and Brockington D (2005) Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries Journal of Politics 67(3): 825-840

Kristofersson D and Navrud S (2005) Validity Tests of Benefit Transfer – Are We Performing the Wrong Tests? Environmental and Resource Economics 30: 279-286

Kristofersson D and Navrud S (2007) Can Use and Non-Use Values be Transferred Across Countries? In: S. Navrud and R. Ready (ed), Environmental Value Transfer: Issues and Methods. Dordrecht, The Netherlands: Springer,

Krosnick J (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5: 213-36

Legget C G, Kleckner N S, Boyle K, Duffield J W and Micthel R C (2003) Social Desirability Bias in Contingent Valuation Surveys Administered Through In-Person Interviews. Land Economics 79(4): 561–575

Li H, Berrens R P, Bohara A K, Jenkins-Smith H C, Silva C L and Weimer D L (2005) Testing for budget constraint effects in a national advisory referendum survey on the Kyoto protocol. J. Agric. Resour. Econ. 30(2): 350-366

Li H, Berrens R P, Bohara A K, Jenkins-Smith H C, Silva C L and Weimer L (2004) Telephone versus Internet samples for a national advisory referendum: are the underlying stated preferences the same? Appl. Econ. Lett. 11(3): 173-176

Lienhoop N and MacMillan D (2007) Contingent Valuation: Comparing Participant Performance in Group-Based Approaches and Personal Interviews. Environmental Values 16: 209-232

Lindhjem H (2007) 20 Years of stated preference valuation of non-timber benefits from Fennoscandian forests: A meta-analysis. Journal of Forest Economics 12(4): 251-277

Lindhjem H and Navrud S (2009) Asking for Individual or Household Willingness to Pay for Environmental Goods: Implication for aggregate welfare measures Environmental and Resource Economics 43(1): 11-29

Lindhjem H and Navrud S (Forthcoming) Using Internet in Stated Preference Surveys: A review and comparison of survey modes. International Review of Environmental and Resource Economics

MacDonald D H, Morrison M, Rose J and Boyle K (2010) Untangling Differences in Values from Internet and Mail Stated Preference Studies. World Congress of Environmental and Resource Economists, Montreal, Canada

MacMillan D, Hanley N and Lienhoop N (2006) Contingent valuation: Environmental polling or preference engine? Ecological Economics 60: 299-307

Maguire K B (2009) Does mode matter? A comparision of telephone, mail, and in-person treatments in contingent valuation surveys. Journal of Environmental Management 90: 3528-3533

Marta-Pedroso C, Freitas H and Domingos T (2007) Testing for the survey mode effect on contingent valuation data quality: a case study of web based versus in-person interviews. Ecological Economics 62: 388-398

McFadden D (1999) Rationality for economists? J. Risk Uncertain. 19(1-3): 73-105

Messonier M L, Bergstrom J C, Cornwell C M, Teasley R J and Cordell H K (2000) Survey Response-Related Biases in Contingent Valuation: Concepts, Remedies, and Empirical Application to Valuing Aquatic Plant Management. American Journal of Agricultural Economics 83: 438-450

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

Nielsen J S (In press) Use of the Internet for willingness-to-pay surveys. A comparison of face-to-face and web-based interviews. Resource and Energy Economics

OECD (2010). *OECD Broadband Portal.* http://www.oecd.org/sti/ict/broadband

OECD (2010). *OECD Factbook.*

Olsen S B (2009) Choosing Between Internet and Mail Survey Modes for Choice Experiment Surveys Considering Non-Market Goods Environmental and Resource Economics 44(4): 591-610

Payne J W, Bettman J R and Schade D A (1999) Measuring Constructed Preferences: Towards a Building Code. Journal of Risk and Uncertainty 19(1-3): 243-270

Roger J L, Howard K I and Vessey J T (1993) Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin 113(3): 553-565

Schulze W, McClelland G, Waldman D and Lazo J H (1996) Sources of bias in contingent valuation. In: D. J. Bjornstad and J. Kahn (ed), The contingent valuation of environmental resources: Methodological resources and research needs. Edward Elgar, Schuman H (1996) The sensitivity of CV outcomes to CV survey methods. In: D. J. Bjornstad and J. Kahn (ed), The contingent valuation of environmental resources: Methodological issues and research needs. Edward Elgar,

Shogren J F (2005) Experimental methods and valuation. In: K.-G. Maler and J. R. Vincent (ed), Handbook of Environmental Economics. . Amsterdam: North Holland,

Stanton J (1998) An empirical assessment of data collection using the internet. Personnel Psychology 51: 709-725

Statistics Norway (SSB) (2010). *The Internet Poll, 2nd Quarter 2009*

Taylor P A, Nelson N M, Grandjean B D, Anatchkova B and Aadland D (2009). *Mode effects and other potential biases in panel-based Internet surveys: Final report.*

Thurston H W (2006) Non-market valuation on the internet. In: A. Alberini and J. Kahn (ed), Handbook on contingent valuation. Edward Elgar,

Tourangeau R, Rips L J and Rasinski K A (2000) The psychology of survey response. Cambridge University Press, Cambridge

van der Heide C M, van den Bergh J C J M, van Ierland E C and Nunes P A L D (2008) Economic valuation of habitat defragmentation: A study of the Veluwe, the Netherlands. Ecological Economics 67: 205-216