Ecological Economics 70 (2011) 1628–1637

Contents lists available at ScienceDirect

Ecological Economics

journal homepage: www.elsevier.com/locate/ecolecon

Analysis

Are Internet surveys an alternative to face-to-face interviews in contingent valuation?

Henrik Lindhjem [a,*], Ståle Navrud [b]

[a] Norwegian Institute for Nature Research, Gaustadalleen 21, N-0349 Oslo, Norway
[b] Department of Economics and Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway

ARTICLE INFO

Article history:
Received 5 December 2010
Received in revised form 23 March 2011
Accepted 1 April 2011
Available online 12 May 2011

JEL classification:
Q51
H41

Keywords:
Internet
Contingent valuation
Interviews
Survey mode
Biodiversity

ABSTRACT

Internet is an increasingly popular data collection mode for stated preference research in environmental economics. However, little is known about how this survey mode may influence data quality and welfare estimates. As part of a national contingent valuation (CV) survey estimating willingness to pay (WTP) for biodiversity protection plans, we assign two groups of respondents either to an Internet or face-to-face (in-home) interview mode. Our design aims to better isolate measurement effects from sample composition effects by drawing both samples from the same sample frame. We find little evidence of social desirability bias in the 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 and equality of mean WTP cannot be rejected. Results are fairly encouraging for the use of Internet in CV as stated preferences do not seem to be significantly different or biased compared to face-to-face interviews.

1. Introduction

The use of stated preference surveys such as contingent valuation (CV) and choice experiments to value benefits of environmental goods is increasing rapidly (Carson, 2011). Although best practice CV studies pay significant attention to questionnaire development and econometric estimation of welfare measures, the choice of data collection mode – mail, face-to-face (f2f), telephone, Internet or a mix – is typically made with comparatively little consideration of its influence on how preferences are formed and stated (Boyle and Bergstrom, 1999). Traditionally, f2f interviewing has been the recommended "gold standard" for CV (Arrow et al., 1993; Mitchell and Carson, 1989), though more research has also been called for to substantiate this recommendation (Schulze et al., 1996). The current trend for conventional "one-shot" CV surveys, however, is to collect data using some form of Internet surveys (e.g. Thurston, 2006). Further, research interest regarding potential impacts of Internet surveying is growing both in the survey methodology and stated preference literatures (Lindhjem and Navrud, 2011).[1] This paper aims to contribute to this research by investigating how Internet surveying may influence stated preferences and derived welfare measures for environmental goods.

The two main sources of potential differences in stated preference results between survey modes are related to methods of sampling (i.e. affecting population 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 aim is to attempt to disentangle these two sources of differences, and assess the possible measurement error due to mode, often termed the "survey mode effect". Are the preferences and the derived welfare measures expressed through the Internet mode different or in practice equivalent to the f2f (in-home) interview mode? In an 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 probability-based pre-recruited panel of survey respondents either to an Internet or a f2f interview mode. This approach keeps sampling better controlled than many previous mode comparison studies. Although the practical recruitment strategy we chose for the survey may have introduced some bias according to mode preference of respondents, we are able to investigate differences between respondents and non-respondents for both modes utilizing background data collected for all members of the survey panel. Both groups receive identical questionnaires administered by a professional survey firm during the same time period to ensure preference stability. We draw on a broad survey methodology literature explaining mode

* Corresponding author. Tel.: +47 98263957.
E-mail address: henrik.lindhjem@nina.no (H. Lindhjem).

[1] Comparing with research on hypothetical bias in CV, List et al. (2004:742) state that "An interesting aspect of CV that has received considerably less attention is whether the survey administration mode is important."

0921-8009/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolecon.2011.04.002

differences especially with reference to social desirability bias and satisficing (shortcutting the response process) and derive hypotheses adapted to the CV context that we investigate in our data.

To our knowledge this is the first well-controlled comparison between Internet and f2f interview modes in stated preference research drawing samples from the same population. Four recent studies compare Internet with f2f interviews in CV of environmental goods (Canavari et al., 2005; Marta-Pedroso et al., 2007; Nielsen, 2011; van der Heide et al., 2008).[2] However, all these studies compare modes with little conceptual guidance about which differences may be expected and why and confound sample effects with measurement effects, as samples are drawn from different populations, in different years, using different questionnaires and survey incentives etc. Two notable exceptions in terms of experimental control, comparing Internet with phone and/or mail rather than f2f, are Taylor et al. (2009) and MacDonald et al. (2010). The general finding of the Internet comparisons and the few studies that have compared other modes in stated preference research (e.g. Maguire, 2009) is that the choice of mode often affects value estimates and other parts of stated preferences, but that the reasons and direction are unclear. 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 discussed in part five that mean WTP from the f2f 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.

2. Survey Mode Effects and CV

2.1. Sources of Survey Mode Effects

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 nature and one of a cognitive nature (Dillman, 2000). The former is related to how cultural norms are invoked differently across modes. The main difference is between a self-administrated survey and the involvement of an interviewer. The most important and well-documented mode effect in this regard is social desirability bias (Groves et al., 2004). The extent of such responding seems to be related to two main factors: the degree of anonymity and trust felt by the respondent. Anonymity is minimised in an f2f interview conducted in the respondent's home, making socially desirable responses more pronounced. On the other hand, interpersonal trust can also emerge between the interviewer and the respondent in a f2f setting resulting in more honest responding. The net effect may be an empirical question. The WTP question can be influenced by social desirability bias since it may be considered a civic virtue contributing to a common good. The effect may depend on the payment format, e.g. whether an open-ended or dichotomous choice (DC) question is used. For the open-ended question with payment card (PC) that we use in our survey, social desirability bias may lead to higher WTP responses. 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 – sometimes used to judge validity of the data – may also be influenced by social desirability bias.

The second factor causing mode differences, the cognitive, is related to individuals' 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.[3] 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). Commonly observed effects include answering "don't know" or refusing (or generally more incomplete responses), selecting the first reasonable response alternative, acquiescence (answering affirmatively regardless of the question), non-differentiation (sticking to the same response category for a sequence of questions), endorsing status quo, random answers (if "don't know" is not an option), and choice of mid-points or only extremes in rating scales. One of the proclaimed advantages of f2f interviews is the motivational effect of the interviewer. This effect may stimulate respondent attention and improve respondent performance. Another advantage is that the interviewer can make it easier for the respondent to understand the questionnaire. These two factors potentially reduce satisficing in f2f interviews compared to Internet surveys. As with social desirability bias, CV questions will be susceptible to satisficing in different ways. In an open-ended WTP question with a PC 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. Due to limited evidence about Internet surveys, it is sometimes assumed to be similar to mail surveys along the normative and cognitive sources of mode effects discussed above (Dillman and Smyth, 2007).

2.2. Comparisons of Internet and f2f Interview Modes for CV

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 and Internet respondents recruited using 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 f2f mode may be an indication of social desirability bias, although there are many confounders, including different sample frames and compositions and a 5% response rate for the Internet survey. There is no consideration of the satisficing issue. In a CV study of WTP for life expectancy gains from reduced air pollution, Nielsen (2011) makes a comparison between Internet and (in-home) f2f interviews recruited from 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. There are few indications that the Internet data has lower validity or is more subject to satisficing strategies. A critical downside of the study is, in addition to

[2] In addition, a few other recent studies either use choice experiments, compare Internet with other modes or contain only brief discussions based on stated preference surveys where mode comparison was not the primary purpose (see e.g. Covey et al., 2010; Li et al., 2009; Olsen, 2009).

[3] Within stated preference valuation, choice experiments may be an exception.

different sample frames, that the two surveys were carried out with more than one year time difference, a period in which preferences may well have changed.

Finally, the study by Taylor et al. (2009) commissioned by US EPA, is 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 lower average welfare measures than the phone survey. They conclude that this result is due to social desirability bias in the phone survey, 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, contrary to common belief in the CV literature and views of the NOAA panel (Arrow et al., 1993) social desirability bias is often found to be larger in telephone than in f2f interviews (see e.g. Jäckle et al. (2010)). In validity checks of the data, Taylor et al. (2009) find no clear indications of lower quality Internet data due to satisficing or other effects, though the unexplained variance in WTP was somewhat higher for the Internet sample. In summary, there is fairly limited evidence of social desirability bias or satisficing specific to f2f CV surveys that have been clearly distinguished from sample effects. A more comprehensive review of the use of Internet and other survey modes in stated preference research can be found in Lindhjem and Navrud (2011).

3. Hypotheses

We propose a few simple indicators adapted from the survey literature that will be tested related to potential satisficing and social desirability effects in the measurement of mean WTP and a few other variables of importance for assessing data quality.

3.1. Satisficing and social desirability

We formulate the following hypotheses:

H1 (satisficing). The share of "don't know" responses to the WTP question is higher for the Internet sample than for the f2f sample.

H2 (satisficing). The distribution of PC responses has lower variance for the Internet than for the f2f sample.

H3 (social desirability). The share of stated zero WTP is higher in the Internet sample than in the f2f sample.

H4 (social desirability). The share of zero respondents that state reasons of protest is higher in the Internet sample than in the f2f sample.

H1 indicates that respondents do not exert the effort necessary to make up an opinion about their WTP and instead choose the "don't know" option. H2 reflects the satisficing tendency often found in survey research to choose mid-points in rating scales (see e.g. Chang and Krosnick, 2009). H3–H4 relate social desirability bias specifically to higher zero and protest responding in the Internet sample.

3.2. Comparison of Mean WTP

H1-4 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 zero responses to the WTP question in the Internet survey reduces mean WTP if the share of these responses due to protesting is the same between samples (as such responses are typically taken out, leaving only the "true" zeros). 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 since 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 f2f sample is unknown. If satificing is highest among low WTP-respondents, which is likely,[4] removing them in the Internet sample will increase mean WTP compared with the f2f sample. Finally, the effect of H2 may go either way for the WTP comparison.

An important question is if the two modes produce welfare measures that for all practical purposes can be considered equivalent. A null hypothesis will often be rejected if sample sizes are large, "resulting in statistically significant differences that are substantively trivial" (Roger et al., 1993:553). For this reason, we also conduct a test of equivalence using a 20% difference level as a starting point, considering other levels for sensitivity. Hence, the following two hypotheses are formulated:

H5a (classic null of no difference). Mean WTP is equal between the Internet and f2f samples.

H5b (non-equivalence of WTP). Mean WTP for the Internet sample is either higher or lower than for the f2f sample by 20% or more.

3.3. Theoretical (construct) Validity

We also compare validity of the data for the two samples in terms of how WTP is related to other variables as predicted by theory. Even if two modes may produce different response distributions for different types of explanatory variables, it is arguably their relationship with WTP that is important:

H6 (construct validity). The relationship between WTP and commonly included explanatory variables is similar between modes in regressions and similar to theoretical expectations.

4. Survey Design and Administration

4.1. 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. The questionnaire was developed following similar forest protection surveys well-tested and tried in the Nordic context (see Lindhjem (2007) and Lindhjem and Navrud (2009)) and adopted to the Internet context following advice e.g. given by Dillman (2000). There are government plans to increase the network of forest reserves to stem the loss of biodiversity (most of which are non-use related such as insects, fungi, mosses and plants). The environmental commodity was therefore specified in the questionnaire as a plans to increase the number of forest reserves from the level of 1.4% of the productive forest area at the time of the survey to 2.8% (a doubling) or 4.5% (minimum recommended by biologists). Respondents received a WTP question for each of these plans in turn, starting with the smallest plan. Responses to the first WTP question are used to compare mean WTP in this study. The text was supplemented with colour maps of current and future forest reserves,[5] and a table giving information about the size and location of new reserves and the likely improvements in the living conditions for main groups of species

[4] See e.g. Holbrook et al. (2003). As education is often correlated with income, and income with WTP, satisficing effects are more likely to be observed among low-WTP respondents.

[5] Six colour photos of (neutral, "non-charismatic") endangered species and forest habitats were also shown as well as pie and bar charts of number and percentage of species in all types of Norwegian habitats.

(see Table A in Appendix A). After the introductory sections, the respondents were reminded of their budget and given a WTP question for their household with the aid of a payment card (PC) for an annual, indefinite earmarked tax increase. The PC contained 24 amounts (ranging from 0 to 15,000 Norwegian Kroner — NOK[6]) arranged on a non-linear scale in a table, including "don't know". The WTP question and the payment card are given in Appendix A. PC was chosen as response format over DC to preserve sample efficiency. According to Boyle's (2003) review it is further far from clear that DC represents the better approach. 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 to supplement panel information. The instrument went through extensive testing in focus groups and pilots using both Internet and interviews.

4.2. Survey Administration in the Two Modes

A randomly recruited panel of 35,000 willing respondents, maintained by the professional survey firm TNS Gallup was used for the survey.[7] To the extent possible in a CV experiment like this, confounding factors not related to survey mode was sought controlled as best as possible (partly based on considerations in Holbrook et al., 2003). First, two groups of respondents were interviewed either in their home or using an Internet survey, 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 for both samples to reduce f2f 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 f2f sample was recruited first by a standard e-mail invitation typically used for all surveys of this type to the panel. It said that the survey (topic of which was not disclosed) would be conducted by f2f 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, to make this as easy as possible for the respondent. 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.[8] Since the panel contains background information about all members, the Internet sample was stratified based on age, gender and education to be as similar as possible to the f2f sample. Fourth, the questionnaire was as identical between modes as practically possible. The Internet survey was a page-by-page (not scrollable) design to make it easy to follow. The f2f interviews were conducted by experienced interviewers of varying age and gender, who were not informed about the purpose of the experiment. Questions were read to the interviewees with the aid of a hand-held pocket computer and answers noted down by the interviewer on the screen. For the most important questions, including the PC, 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. Identical maps, colour photographs and graphs were displayed from an interview folder in the same order as in the Internet survey.

As a probe of social desirability bias, we asked interviewers to openly assess after the interview to what degree they thought the situation pressurized respondents in any way. 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 min, while completion times for the Internet survey were somewhat shorter, at around 25–30 min. This difference is primarily because it takes longer to read and display information, maps, cards etc. in the interview than it does for the respondent to click through the Internet survey. However, the time may be shorter also because of satisficing. As indicator of respondent effort, we also measured the time it took Internet respondents to read and answer three different parts of the survey. However, since we unfortunately did not have this information for both modes, we can only utilize it in the analysis of the Internet results. Fifth, the surveys were conducted during the same period in October and November 2007 to ensure preference stability between modes. Finally, the same token incentive payment to reply were given to both samples avoiding any related selection bias, and all respondents were interviewed individually.

5. Results and Analysis

5.1. Samples and Response Rates

668 respondents first accepted to be interviewed f2f, 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% final-stage response rate. The original number of e-mail invitations for the f2f interviews was not given by the survey firm, precluding calculation of the more appropriate multi-stage response rate. For the Internet survey, a random sample of 645 was contacted, of which 385 completed the survey. This is a response rate of 59.7% for the Internet survey. 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, and both of these are fairly good.

The socio-economic characteristics of the two modes, both for gross and respondent (net) samples (i.e. those from the gross sample who responded), are given in Table 1. All information (except for average household income) is taken from the panel database updated in the same year as the survey. Between the gross samples there are no statistical differences between age and gender, but there are some differences between income and education distributions at the 10% 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 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.

Further, the rate of Internet use is not different between samples, i.e. there is no tendency of self selection in this regard. The comparison of observable characteristics indicates 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. However, we cannot rule out that some respondents have been self-selected due to their mode preference, even though we cannot detect this based on observable characteristics. We proceed by testing our hypotheses and investigate validity of the data without

[6] 1 Norwegian Krone (NOK) = ca 0.125 Euro at time of study.

[7] TNS Gallup uses no form of self recruitment (such as website advertisements etc.), which is a common form of Internet survey recruitment, including in panels such as Harris Interactive in the US.

[8] 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 check the effect of this along observable characteristics on the gross and net samples in the next section.

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

Socio-economic variables	Gross samples			Respondent (net) samples		
	F2f (n = 398)	Internet (n = 645)	Test stat. between modes	F2f (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			$\chi^2 = 8.4^{*}$			$\chi^2 = 5.8$
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 indicated in more detailed income categories than for the income information in the gross samples. Pearson's chi-square test used to compare frequency distributions.

weighing the samples by socio-economic characteristics or conducting further investigation of non-response effects.[9]

5.2. Satisficing and Social Desirability

We start by reporting the results from the satisficing hypotheses (H1 and H2). The data rejects H1: 11% of Internet respondents and 8% of the f2f respondents state "don't know", a difference in the expected direction, though not significant on the 10% level (see row three in Table 2). H2 of lower variance for the WTP distribution from the Internet sample is also rejected using a likelihood ratio test for the parametric WTP model explained in the next section (row four in Table 2).

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 s, while median times to complete the two sections on current policies and new policies including the WTP question were 105 s 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.[10] This means that the less time respondents spend on the WTP question, the more likely they are answering "don't know" or zero. This may be an indication that both these response types result from satisficing strategies.[11]

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 at 19.3 and 18.9%, respectively (see row five in Table 2). Hence, there is no evidence that the f2f situation makes it socially harder for respondents to state a zero response. Second, we tested whether two types of protesting which slightly different interpretations for social desirability were more common in the Internet survey (H2). When answering zero respondents were asked in standard CV fashion to state up to two reasons from a list to enable identification of protest responses. A strict interpretation of protest would be to include all those who state zero even if the good has a positive value to them and they are not prevented from paying by an income constraint. Using this interpretation leaves shares of protest that are not statistically significant (90.65 vs. 88.06%, see first row below H4 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 higher perceived "social punishment" in the f2f situation, e.g. related to taxes and responsibility for causing or solving the problem,

Table 2
Test results for indicators of satisficing and social desirability bias.

Hypotheses: satisficing and social desirability		Survey modes		Mode comparison	
		F2f (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	$\chi^2 = 14.27$[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

[a] Likelihood-ratio test of equality of standard error, sigma (σ), as explained in footnote 12.

[9] A more comprehensive analysis could have included both running a Heckman sample selection model (Heckman, 1979) or weighing samples by demographics.

[10] For "don't know"responses the result is robust at the 10% level for times from 0–600 seconds (i.e. 10 min), which includes 95.6% of responses. For zero responses the result is robust at the 5% level for times from 0 to 4000 s (67 min), which includes 98.7% of responses.

[11] Some zero respondents may be sure that they do not want to pay, and therefore do not need much time to state zero.

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

Hypothesis		Mean WTP F2f: (95% CI)	Mean WTP Internet: (95% CI)	Mode Comparison result (p<0.1)
H5a	Equality of mean WTP	1819 (1539, 2100)[a]	1566 (1261, 1871)[a]	Non-rejection
	N	218	269	

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

were distinguished from idealistic reasons or response difficulties. 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% protests in the f2f sample and 70.90 in the Internet sample (last row in Table 2). Overall then, little evidence has been found in our data for the hypotheses of social desirability bias and lower level of satisficing in the interviews.

5.3. Comparison of Mean WTP

To test H5a and 5b, 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) (se also the WTP distribution from the PC in Fig. B in Appendix B). Since the stated WTP amounts have a skewed distribution with the familiar long right tail, a log-transformation of WTP was applied.[12] 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. Mean WTP is given in Table 3.

The mean for the f2f sample is somewhat higher at NOK 1819 than the NOK 1566 for the Internet sample. We calculate 95% confidence intervals around the respective means based on a bootstrap (10,000 draws with replacement) from each of the sample distributions. Since the confidence intervals are overlapping we cannot reject H5a that mean WTP are equal between modes on the 5% level (see bootstrap distributions in Fig. 1).

We also investigate whether the difference between means is of practical importance — (H5b). To test this hypothesis we combine the two bootstrapped mean WTP distributions in Fig. 1 into a single distribution of the differences in mean WTP for the two modes (see Fig. 2). First, since only 87.95% of the distribution is larger than zero we cannot reject H5a at the 10% level.

We test how much of the distribution is outside different equivalence intervals (see results in Table B in Appendix B). First, testing whether mean WTP for the Internet sample is higher or lower than 20% of mean WTP for the interview sample (i.e. ±NOK 364) leads to non-rejection since around 30% of the distribution is contained outside this bound. The cut-off point between rejection and non-rejection is 28% difference, at the 10% confidence level (row five in Table B). If we keep to the 20% equivalence level, we are unable to reject any of our hypotheses H5a or H5b. This means we cannot conclude "either a sizeable difference or a reliably small difference" (Roger et al., 1993: 563) between modes. However, the sensitivity analysis shows that increasing the acceptable level of difference to 30% would comfortably reject H5b.

[12] 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.

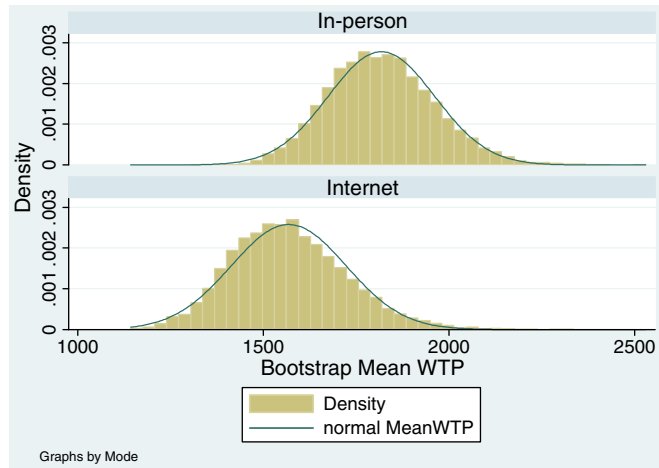

Fig. 1. Distribution of bootstrapped mean WTP from the two samples (10,000 draws with replacement).

5.4. Theoretical Validity

Table 4 presents results of four double log interval regression models, investigating theoretical validity between modes. Log of WTP is the dependent variable. 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.

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 state 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 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. 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 (and similar to regression results typically observed in the CV literature).

To complement the analysis of social desirability and satisficing above, we included some additional variables. First, whether

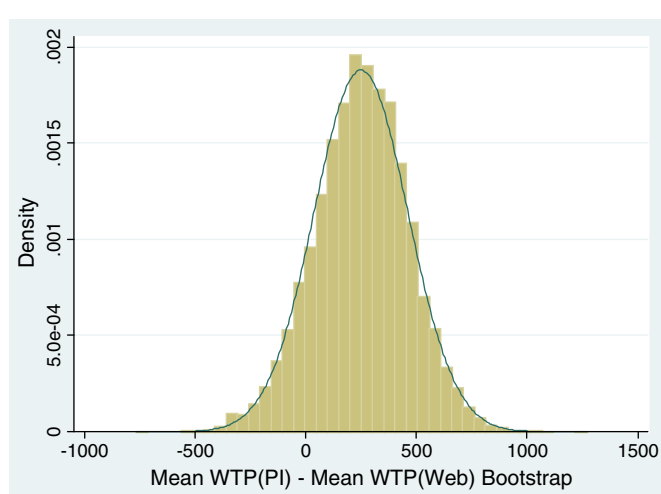

Fig. 2. Distribution of bootstrapped mean WTP(interview) — mean WTP(Internet).

Table 4
Estimation results for face-to-face and Internet modes.

Independent variables		F2f 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 and 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 PC. Ln: means log transformations. Regression in STATA Version 9.2.
[a] Variable information taken from respondent panel database.
[b] Time use information only available from Internet survey.
[c] A few respondents did not state income, so these observations have been excluded.

respondents indicated that they thought it was "very hard" to answer the WTP question is included as a dummy variable, "Difficult". Interestingly, respondent difficulty seems to translate into significantly lower WTP only in the Internet mode. This result may 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. None of these coefficients are significant, indicating fairly consistent interviewing (see Model 2). The indication from interviewers if they thought respondents felt pressured (variable "IntPress") had no effect on WTP. Finally, as noted, we measured the time it took Internet respondents to complete three separate sections, included as variables, "Time1", "Time2" and "Time3" (where the third is completing the WTP question). 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 (see Model 4). This cannot really be explained by the fact that interested respondents may spend more time on the survey and state higher WTP, since time spent on the other parts of the survey has no effect on WTP.

6. Discussion and Conclusion

In a CV experiment we have conducted the first well-controlled study of whether stated preferences are different between Internet or in-home f2f interview survey modes. Checking in particular for indications of social desirability bias and shortcutting of the response process (satisficing), 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 — PC) is similar between modes. There is also no tendency of PC responses being more closely clustered together in the Internet mode. Mean WTP is somewhat higher in the f2f sample, though we cannot reject that mean WTP in the two modes are equal on the 10% level. Considering equivalence, we can reject that the WTP difference is more than 30%, but fail to reject an equivalency bound of 20% on the 10% confidence level. While 20–40% has been suggested as acceptable for benefit transfer settings (Kristofersson and Navrud, 2007), it may be argued that levels in the lower end of this interval could be applicable when comparing survey modes. The agreed-upon standard adopted in for example pharmaceutical research for equivalence of population means is ±20% (Roger et al., 1993). Equivalency considerations is important as the current growth in the use of Internet in stated preference research will enable large, low cost split-sample tests that may typically find significant, though often practically trivial, differences between treatments. Finally, we find that the two survey modes show the same degree of construct validity for different WTP model regressions.

So, can our results be generalized to the broader population and to people who are not survey panel members? Comparing our samples from Oslo with the whole survey panel for the country indicates that our samples are quite similar to the general survey panel population in terms of age, gender and employment rate. However, people with more than four years university education are somewhat

overrepresented (around 20% compared with 13% in the panel). This mirrors the higher education levels in urban areas. We also see this overrepresentation to some extent when comparing the overall survey panel with the general population. It is not clear how this would play out in terms of satisficing and social desirability and measurement effects more generally. We do, however, find that lower education respondents tend to state "don't know" more often, and that they do so to a larger extent in the Internet mode than in the f2f mode (33.3% vs. 26.3%). We also found, as shown in Table 4, that perceived survey difficulty (as may also be more common among low-education respondents) tended to give lower WTP only in the Internet sample. These two combined effects may result in lower WTP from Internet surveys when conducted for a more representative sample with a higher share of low-education respondents. These effects may be less pronounced for surveys attempting to value less complex goods that are more familiar to people than forest biodiversity.

In terms of age, there is a similar underrepresentation of the higher age groups in our study (around 9% in >60 years group in our samples compared to 24% in the general population). The degree to which older people have response patterns that would influence measurement errors differently than for younger people has not been investigated in our study, and would be a topic for further research. It is also worth noting that Internet surveys currently tend to have lower response rates than other survey modes, which may introduce stronger elements of self-selection and non-response biases compared to e.g. f2f surveys when generalising results (Manfreda et al., 2008). Another concern voiced about generalising from Internet panels is the potential time-in-sample effects.[13] We did not have information from the survey company about the survey history of our respondents, though TNS Gallup ensured that their panel was regularly changed. Given the high familiarity and frequency of Internet use in Norway[14] (e.g. all tax and most other government forms are now Internet-based), we also find it unlikely that panel respondents are substantially better able to answer Internet surveys than the general population.

Further, we are also cautious of generalisation to other goods and survey types, 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, unlike Norway, where it is not considered "polite" to disagree (see e.g. Ehmke et al. (2008) for an international comparison of hypothetical bias). Given the lack of evidence of survey mode effects in our study, Internet surveys may well provide a reliable, low-cost alternative to f2f interviews. However, continued use of f2f interviews may also be seen as a compromise between mail, phone and Internet and the more deliberative approaches recently introduced in CV to facilitate better learning or construction of preferences for complex and unfamiliar goods (see e.g. MacMillan et al. (2006)). In any case, for reasons of cost, convenience and opportunities for better designs and experimentation, Internet either as stand-alone applications or as the primary mode in mixed-mode surveys, is set to grow tremendously. Whereas the coverage and representativeness concerns about Internet are likely gradually to be reduced in Western countries, potential measurement differences between modes will remain. In this respect, our results are quite encouraging in that stated preferences from the Internet survey seem not to be significantly different or less reliable compared to interviews. More well-controlled mode comparisons is, however, necessary not only to document mode effects, but to better pin down and understand their causes. This would help us better controlling potential measurement biases within an acceptable range in future CV applications.

[13] Changes in responses due to the experience of having previously being surveyed.
[14] According to Statistics Norway (2010), in the second half of 2010, 90% of the population had access to the Internet, of which 83% is broadband.

Acknowledgements

This paper is based on surveys funded by project no. 173230 of the Research Council of Norway. We are grateful for comments from Olvar Bergland, Bob Berrens, Kevin Boyle, Mark Dickie, Shelby Gerking, Karine Nyborg, Randall Rosenberger and V. Kerry Smith. The usual disclaimer applies.

Appendix A

Table A
Information about the two alternative protection plans in the CV survey[a].

Today's situation: 1.4% of productive forest area	Increase to 2.8% of productive forest area	Increase to 4.5% of productive forest area
Total protected area: 1.05 million mål[b]	Total protected area: 2.1 million mål[b]	Total protected area: 3.3 million mål[b]
No new reserves	Twice today's protected area, 2/3 in Eastern Norway	Three times today's protected area, 2/3 in Eastern Norway.
Types of species: Same types of species as today remain vulnerable or threatened	Types of species: Large improvement for lichens, mosses, plants and fungi	Types of species: Very large improvement for lichens, mosses, plants and fungi. Large improvement for insects.

[a] WTP question for 2.8% plan used in our survey mode comparison.
[b] A Norwegian measure of area. 10 mål equal one hectare.

WTP question: "We ask you first to consider the smallest of the two protection plans, the one for 2.8%. Then you will be asked about the other plan. What is the most your household almost certainly is willing to pay in an extra annual tax earmarked to a public fund for increased forest protection from today's level of 1.4% to 2.8% of productive forest area?

Choose the highest amount, if anything, your household almost certainly will pay."

(Check one)
☐ 0
☐ 25
☐ 50
☐ 100
☐ 300
☐ 500
☐ 700
☐ 900
☐ 1100
☐ 1400
☐ 1800
☐ 2200
☐ 2700
☐ 3200
☐ 3800
☐ 4400
☐ 5100
☐ 5800
☐ 7000
☐ 8500
☐ 10,000
☐ 13,000
☐ 15,000
☐ More than 15,000
☐ Don't know

Appendix B

D -

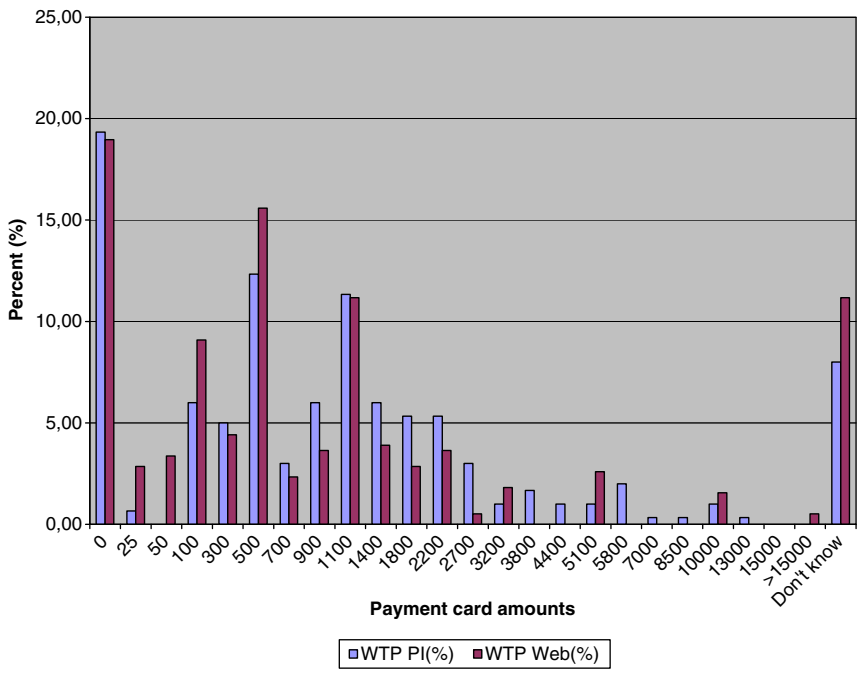

Fig. B. Household WTP distribution as indicated in payment card. Norwegian Kroner, annual amounts for an indefinite period.

Table B
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	Cannot reject
	Non-equivalence, 20%	±364	30.17	Cannot reject
	Non-equivalence, 25%	±455	16.04	Cannot reject
	Non-equivalence, 28%[a]	±511	9.99	Rejection
	Non-equivalence, 30%	±546	7.55	Rejection
	Non-equivalence, 40%	±728	1.10	Rejection

[a] 28% is the difference between means, which allows rejection at the exact 10% level.

References

Arrow, K.J., Solow, R., Leamer, E., Portney, P., Radner, R., Schuman, H., 1993. Report of the NOAA panel on contingent valuation. Federal Register 58, 4601–4614.

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

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

Cameron, T.A., 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., 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.

Carson, R., 2011. Contingent valuation: a comprehensive bibliography and history. Edward Elgar.

Chang, L., 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.

Covey, J., Robinson, A., Jones-Lee, M., Loomes, G., 2010. Responsibility, scale and the valuation of rail safety. Journal of Risk and Uncertainty 40 (1), 85–108.

Dillman, D., 2000. Mail and Internet Surveys: The Tailored Design Method. John Wiley & Sons.

illman, D.A., 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., List, J.A., 2008. Is hypothetical bias a universal phenomenon? A multinational investigation. Land Economics 84 (3), 489–500.

Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2004. Survey Methodology. Wiley.

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

Holbrook, A.L., Green, M.C., 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.

Jäckle, A., Roberts, C., Lynn, P., 2010. Assessing the effect of data collection mode on measurement. International Statistical Review 78 (1), 3–20.

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

Krosnick, J., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5 (3), 213–236.

Li, H., Jenkins-Smith, H.C., Silva, C.L., Berrens, R.P., Herron, K.G., 2009. Public support for reducing US reliance on fossil fuels: investigating household willingness-to-pay for energy research and development. Ecological Economics 68 (3), 731–742.

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., 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., Navrud, S., 2011. Using Internet in Stated Preference Surveys: A review and comparison of survey modes. Submitted to International Review of Environmental and Resource Economics.

List, J.A., Bohara, A.K., Kerkvliet, J., 2004. Examining the role of social isolation on stated preferences. The American Economic Review 94 (3), 741–752.

MacDonald, D.H., Morrison, M., Rose, J., 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., Lienhoop, N., 2006. Contingent valuation: environmental polling or preference engine? Ecological Economics 60, 299–307.

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

Manfreda, K.L., Bosniak, M., Berzelak, J., Haas, I., Vehovar, V., 2008. Web surveys versus other survey modes — a meta-analysis comparing response rates. International Journal of Market Research 50 (1), 79–104.

Marta-Pedroso, C., Freitas, H., 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 (3-4), 388–398.

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

Nielsen, J.S., 2011. Use of the Internet for willingness-to-pay surveys. A comparison of face-to-face and web-based interviews. Resource and Energy Economics 33 (1), 119–129.
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.
Roger, J.L., Howard, K.I., 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., Lazo, J.H., 1996. Sources of bias in contingent valuation. In: Bjornstad, D.J., Kahn, J. (Eds.), The contingent valuation of environmental resources: methodological resources and research needs. Edward Elgar.
Statistics Norway (SSB), 2010. The Internet Poll, 2nd Quarter.
Taylor, P.A., Nelson, N.M., Grandjean, B.D., Anatchkova, B., Aadland, D., 2009. Mode effects and other potential biases in panel-based Internet surveys: final report. US EPA.
Thurston, H.W., 2006. Non-market valuation on the Internet. In: Alberini, A., Kahn, J. (Eds.), Handbook on contingent valuation. Edward Elgar.
Tourangeau, R., Rips, L.J., Rasinski, K.A., 2000. The Psychology of Survey Response. Cambridge University Press.
van der Heide, C.M., van den Bergh, J.C.J.M., van Ierland, E.C., Nunes, P.A.L.D., 2008. Economic valuation of habitat defragmentation: a study of the Veluwe, the Netherlands. Ecological Economics 67 (2), 205–216.