Internet CV surveys – a cheap, fast way to get large samples of biased values?

Henrik Lindhjem[ab*] and Ståle Navrud[a]

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

[b] ECON Pöyry, P.O. Box 5, N-0051, Oslo, Norway

[*] Corresponding author. E-mail: henrik.lindhjem@umb.no.

Internet CV surveys – a cheap, fast way to get large samples of biased values?

Abstract

With the current growth in broadband penetration, Internet is likely to be the data collection mode of choice for contingent valuation (CV) and 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 controlled field experiment as part of a large national 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 interview mode. Our design is better able than previous mode comparison studies to isolate measurement effects from sample composition effects. Looking in particular for indications of social desirability bias and satisficing (shortcutting the response process) 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 very similar between modes. Mean WTP is somewhat higher in the interview sample, though we cannot reject equality on the 10 percent level. We also consider equivalence, i.e. whether the WTP difference is larger than a practically trivial predetermined bound. We can reject that the difference is larger than 30 percent, but fail to reject an equivalency bound of 20 percent on the 10 percent level. Results are quite encouraging for the use of Internet as values do not seem to be significantly different or substantially biased compared to in-person interviews.

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

JEL Classification: Q51, H41

Introduction

One way the economics profession tries to defend 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 published 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. Granted, current best practice CV studies do pay significant attention to questionnaire development and testing to ensure that survey instruments work as intended. However, the choice of data collection mode – mail, in-person, telephone, Internet[1] or a mix – is typically made with comparatively little evidence (or consideration) of its influence on how preferences are formed, stated and added up. 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 question is asked), rather than

[1] 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. Different types of Internet samples are discussed subsequently.

merely revealed or uncovered by it[2]. Traditionally, in-person interviews has been the recommended “gold standard” for CV (Mitchell and Carson 1989; Arrow et al. 1993) for informational and response rate advantages, though it has weaknesses of its own compared to other modes (e.g. interviewer effects, high cost). 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 very low costs. Judging from the current growth in broadband and Internet 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))[3]. 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. Arlinghaus and Mehner 2003; Tsuge

[2] This has been a rather 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). This “constructive” 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.

[3] An average of 18.9 per cent of OECD inhabitants had some kind of broadband in June 2007, up from only 3.8 percent in 2002 (OECD 2008). In Norway, the place of this study, 64 percent had broadband access in the last quarter of 2007, up by 17 percent in one year. Internet access (incl. non-broadband connections) for 2007 was 83 percent (Statistics Norway 2008). 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).

and Washida 2003; Berrens et al. 2004; Hoehn et al. 2004; Schwappach and Strasman 2006; Thurston 2006; Damschroder et al. 2007; Lindhjem and Navrud 2008). 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 the "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 or are the two modes equivalent? Which mode differences can be expected? How can observed 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 large 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 willing survey respondents either to an Internet or an in-person interview mode. We can thus control the effects of sample composition and measurement errors due to mode differences[4]. Both groups get 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 modes and considerations of equivalence, i.e.

[4] 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".

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 to test whether two drugs have equivalent properties (see e.g. Welling et al. 1991)[5], and has increasingly been introduced in the social sciences: in psychology (Roger et al. 1993), survey mode research (Stanton 1998; Epstein et al. 2001) and in benefit transfer in environmental economics (Kristofersson and Navrud 2005; Johnston 2007).

To our knowledge this is the first controlled comparison between Internet and any other mode in stated preference research drawing samples from the same population. Three other studies compare Internet with in-person (on-site) interviews (CV) (Marta-Pedroso et al. 2007), with mail survey (choice experiment) (Olsen 2007), or with telephone recruited computer assisted survey (CV) (Dickie et al. 2007). In addition, Banzhaf et al. (2006) (mail), Berrens et al. (2003) (telephone), Hynes and Hanley (2006) (in-person on-site) contain brief Internet comparisons. The studies to date have compared modes with little conceptual guidance about which differences may be expected and why, and typically confound sample effects with measurement effects. The general finding of the Internet comparisons, and the few that have compared other modes than Internet in CV, is that the choice of mode do affect value estimates and other parts of stated preferences, but that the reasons and direction are unclear (as also observed by Boyle (2003))[6]. We start in the next section by reviewing the theory and

[5] 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).

[6] A number of meta-analyses of the environmental valuation literature also document systematic differences in welfare estimates depending on survey modes (see e.g. Lindhjem (2007), Rosenberger and Loomis (2000) and Johnston et al. (2005)).

evidence of mode effects in survey research and CV of relevance to our mode comparison. 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 in the in-person interview sample is somewhat higher than in the Internet sample, though not significantly so at the 10 percent level. However, even if we cannot reject the traditional null hypothesis, mean WTP may still not be considered equivalent – depending on the level of difference considered acceptable. 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

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 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)[7]. The

[7] It is worth noting that the NOAA panel made recommendations for natural resource damage assessments for use in e.g. court cases as basis for compensation payments. As such the guidelines are arguably stricter than required for CV research more generally (see e.g. Navrud and Pruckner (1997)).

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 the mail survey guru, Don A. 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[8], 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 nature and one of a cognitive or psychological nature (Dillman 2000). The former, the normative factor, is related to how cultural norms in some way are invoked differently across modes leading to culturally constrained responses. The main difference is between a self-administered situation and the

[8] 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. Edwards and Anderson (1987) (telephone and mail), Messonier et al. (2000) (mail and in-person), Harpman et al.(2004) (mail), Hudson et al. (2004) (phone, mail & Internet) for a discussion of such effects in CV.

involvement of an interviewer. In addition, there may be (smaller) differences between mail and Internet on the one hand and telephone and in-person interviews on the other. The most important, and well-documented mode effect in this regard, is according to Groves et al. (2004), social desirability bias[9]. The latter factor, the psychological, is individuals' cognitive processing of information and questions in particular how (aural and/or visual) stimulus produces different responses across modes. Since little is actually known about the Internet as a survey tool it is sometimes assumed to be similar to mail surveys along the two main dimensions (Dillman and Smyth 2007). First, we discuss the two main mode effects related to the normative and psychological factors, before we review results of the limited CV literature in this area.

Social desirability bias

People like to appear favourably in the eyes of others as well as in their own. Thus a socially desirable response can either be an intentional lie (or less strong: "polishing" of the truth or "response edit") or sometimes self-deception. Further, the response may be retrieved and then deliberatively edited after exerting much effort (Holtgraves 2004), or be a result of shortcutting the response process (see below) and merely echoing what is thought socially desirable or politically correct. 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 while answering the survey. Social distance is minimised in an in-person interview conducted in the respondent's home. The cost for the respondent in terms of fear of frowns of disapproval or other signs of disrespect from the interviewer upon a perceived socially undesirable response is therefore the highest. Even if the respondent is allowed to submit a response anonymously (e.g. on a note put in a "ballot box", as suggested

[9] The tendency of respondents to give the answer they feel they ought to give is sometimes more generally termed ”compliance bias” (Green and Tunstall 1999).

by the NOAA panel) the social desirability effect is unlikely to go away as the respondent may still be under the spell of a "focusing illusion" related to issue at hand (Schkade and Kahneman 1998) or get a (slightly) troubled conscience. The cost of an honest, but socially undesirable response is the lowest answering mail and Internet surveys, while telephone occupies a middle position (channels of unfavourable reactions from interviewer are more limited). On the other hand, a great deal of interpersonal trust can emerge between an interviewer and the respondent in a face-to-face interview, especially in the respondent's home[10]. This may both put to rest respondent concerns about whether responses will be misused, go astray or be linked to her identity, and make the respondent open up and be more honest resulting in less socially desirable responding. Concerns over anonymity are likely to be stronger in Internet, phone and mail surveys. Internet may embody an additional fear of anonymity breach compared to mail and telephone, due to well-known cases of identity thefts ("phishing"), hacker break-ins etc. (in addition to the general fear of new technologies). In a comparison between telephone and in-person surveys Holbrook et al. (2003) argue that the effects of social distance and interpersonal trust on social desirability bias may cancel each other out empirically. Contrary to common beliefs, and those held by the NOAA panel, socially desirability bias is often found to be larger in telephone than in in-person interviews, at least for sensitive questions (e.g. questions of race, alcohol use etc.) (see e.g. Groves et al. (2004) or Jäckle et al. (2006))[11]. In addition to social distance and rapport with an interviewer, there may conceivably also be other cues that can influence whether a respondent will answer in a socially desirable way in other modes. For example, it is possible that respondents to a

[10] Interviews on-site or in other public settings (e.g. in shopping malls) may feel too rushed to achieve the same level of rapport and may also put limitations on the confidentiality of the interview if there are other people at the site. However, little is known about effects of different types of in-person interviews.

[11] The survey literature has also documented other interviewer effects that may or may not indicate social desirability bias e.g. related to the origin, skin colour, sex or dress of the interviewer (see e.g. Groves et al (2004)).

larger degree will try to “satisfy” the administrators of an Internet survey about a salient common good such as research, environmental protection, public health or similar, than surveys of a more “neutral” or commercial nature.

The relative importance of the different effects related to social desirability discussed above is hard to judge for CV. First, it is clear that 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 (see next section). 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 (Blamey et al. 1999), in in-person or telephone interviews than in Internet or mail modes[12]. 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[13]. 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

[12] A special case of yea saying is ”warm glow”, in which respondents value giving per se (Andreoni 1990). Warm glow is also likely to be more pronounced in interviews than in self-administered modes.

[13] 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)).

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[14] 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.

Satisficing

To execute the response process well, respondents need to exert some degree of effort and in CV generally more so than in other surveys (Mitchell and Carson 1989). 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). Which level of effort is sufficient for an optimal response – and therefore the degree of satisficing – depends on a combination of task difficulty and respondent ability and motivation. Ability is often proxied fairly accurately by education level. In the language of economics, respondents are likely (to behave as if) conducting a constrained optimisation, which in most cases will lead to a response below the global optimum. Surprisingly little economic research has been conducted to better understand the way humans process complex information in CV surveys (and in other choice contexts) and allocate mental effort resources to this task, even though CV researchers for some time have studied impacts on WTP (and other response variables) of varying the quality and quantity of information (and various stimuli such as colour

[14] Income is sometimes not reported at all typically forcing CV analysts to exclude such observations from the sample.

photographs etc.) (Blomquist and Whitehead 1998)[15]. Promising recent research explicitly studying complexity, information processing and effort allocation include Berrens et al. (2004) and DeShazo and Fermo (2002) in a CV context, while Gabaix et al. (2003) provides a more general theoretical framework. Time-strapped, unmotivated respondents' 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), 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. These measurement errors are sometimes difficult to separate from socially desirable responding[16] and response order effects not related to satisficing (Groves et al. 2004). Measurement errors due to satisficing or for other reasons have not been much studied in Internet surveys (Dillmann and Smyth 2007), though some recent studies document similarities with typical mail survey errors (see e.g. Tourangeau et al. (2004; 2007)).[17]

[15] As observed by Smith (2000:363), taking the long view on environmental economics research: "Choices that are informative about an individual's preferences are difficult ones for that person to make. Under these conditions it seems that survey approaches must address the factors that influence how much effort people will expend to understand "proposed" choices. [..] Research to date has not provided a model to describe how the choice context and question format influence respondent's willingness to exert effort to understand the full dimensions of choices when there are no tangible financial incentives."

[16] However, Holtgraves (2004) found that socially desirable responding was related to longer response times, indicating that such responding may be more common as a deliberate editing effort, rather than a as a result of satisficing.

[17] However, one important difference between mail and Internet is that the questionnaire may not be displayed in the same way on all computer screens (i.e. due to screen settings or browser software etc.) making it harder to control effects (Dillman and Smyth 2007).

The main point here is how modes affect the tendency to satisfice for different types of questions in a CV survey. 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[18]. These two factors reduce respondent benefits of satisficing in interviews compared to the Internet mode. On the other hand an in-person interview may also carry costs in terms of time and pressure put on the respondent to answer, inducing satisficing. Internet surveys may also carry a (fast depreciating) novelty benefit, can be easier to understand than a mail survey (e.g. because respondents are automatically directed to the next question through filters, pictures and illustrations can be provided more easily etc.), and the respondent can answer in her own time. The net effect for Internet and in-person modes may be difficult to assess for CV, although it is generally agreed that satisficing may be a bigger problem in self-administered than interview surveys

[18] Answers to questions respondents may have in CV surveys are typically written down for interviewers to read consistently if asked. Text may also be read a second time. No extra explanation is normally given to increase the understanding of respondents. This is called "standardized interviewing" giving high priority to replicability of scientific findings. However, standardized interviewing is controversial in survey research. Opponents argue that exposing people to the same words does not mean they are understood in the same way, and that it is an unnatural form of interaction that is particularly inappropriate when the interviewer can clearly see that the respondent is misunderstanding (see discussion in Chapter 9.6 of Groves et al. (2004)).

(Holbrook et al. 2003)[19]. 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.

Mode effects 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. The only other study that compares in-person interviews (on-site) and Internet modes we are aware of is Marta-Pedroso et al. (2007). Sampling 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), though no statistical test was conducted. The higher mean WTP in the in-person mode is 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, to draw such a conclusion firmly (which the authors sensibly also do not do). There is also no consideration of the satisficing issue in the study. As such it is more a practical comparison of modes than a controlled experiment. In a choice experiment setting Olsen (2007) investigated preferences for protecting recreational use values from motorway encroachment in two municipalities in Denmark comparing a pre-recruited Internet

[19] Apart from satisficing, measurement biases between modes may also be due to other reasons e.g. in the way the aural vs. visual senses are stimulated, resulting in different processes through which the meaning of a question and the response alternatives are comprehended.

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 an indication of real response differences. Comparing mean WTP between samples Olsen (2007) concludes that it cannot be rejected that preferences from the two modes are identical. He then draws the, in our view, somewhat premature (and unsupported) 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. The authors speculate that the lower quality may be due to Internet respondents being more distracted (both by family members, TV and by having the option to leave and complete the survey at a later time) or panel attrition (time-in-sample effects)[20]. Higher motivation among the RDD respondents accepting to travel to a University campus for little compensation to complete the survey, as pointed out by the authors as a possible reason, may however in our view, be the most likely reason. 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). Banzhaf et al. (2006) conduct a high-budget CV survey of the WTP for ecological improvement in the Adirondack Park in the USA. Two thirds of the sample was Internet panellists and one third mail respondents (from two different sample

[20] Changes in responses due to the experience of having previously being surveyed.

frames). Their main focus is not on mode effects, but they conduct a brief convergent validity check and cannot reject the hypothesis of equal WTP from the two samples weighted for differing demographics. Mainly addressing the issue of representativeness of two types of Internet samples[21] compared with a RDD telephone sample for political research, Berrens et al. (2003) also assess questions of environmental attitudes and WTP.[22] 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 with bid price). Finally, of the studies including one Internet sample in their comparison, Hynes and Hanley (2006) estimate the demand for kayaking in Ireland using a travel cost survey administered in-person on-site and to (self-selected) Internet respondents following a survey link on a kayaking enthusiast webpage. With the caveats of self-selection and small samples, they find that the two samples could generally be pooled. No survey mode effects related to social desirability or satisficing were reported (and such effects may also work differently in a travel cost context).

A few other studies investigate in-person interview modes in environmental valuation, some in comparison with telephone or mail. List et al. (2004) find in a face-to-face field experiment of students that the share of both stated and actual WTP for the establishment of an environmental policy research centre is reduced when the degree of social isolation or

[21] 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).

[22] The few recent studies comparing Internet with other modes in political and other social science research find similar results between modes for the constructs of interest (see e.g. Fricker et al. 2005; Denscombe 2006; Malhotra and Krosnick 2007; Sanders et al. 2007).

anonymity is increased. It is uncertain how this result can be transferred to a more general CV mode context, but it confirms that social desirability bias may be a problem in in-person surveys. Legget et al. (2003) is a more traditional CV study testing whether the NOAA-panel recommendation to allow in-person respondents to submit their WTP bid in a "ballot box" will reduce social desirability bias equivalent to an anonymous mail-back option. Surveying visitors to a national monument on-site, they find using an open-ended PC approach that mean WTP was approximately 23 percent higher for the interview than for the mail-back option. Though there may have been some small self selection problems (e.g. the ones agreeing to an interview may view the issue more favourably than the mail respondents), it is still an indication that the social pressure felt by the respondent is carried over in the statement of WTP, even if anonymous. The overall validity of the data in the two modes or degree of satisficing was not considered. In a transition country context Davis (2004) compares hypothetical WTP for water service improvements across four modes recruited in a practical, rather than controlled way: household and intercept in-person modes, focus groups and telephone. She finds that mean WTP is between 23 and 78 percent higher in the telephone survey compared to the three other modes (between which there are small differences). This may be an indication that social desirability bias is higher in telephone than in-person interviews (as discussed above). Further, she finds that "don't know" or "not sure" responses to the WTP question, indications of satisficing, are 2-3 times higher for the telephone sample than for the two in-person modes. Davis (2004) speculates that this finding is due to time pressure on the telephone (though she also acknowledges that sample compositions may be important in explaining differences). Maguire et al. (2002) find lower WTP in in-person interviews than in mail and telephone surveys, though the confounding of different sample frames limits the usefulness of these results. Finally, for the studies comparing in-person with other modes Hanley (1989) finds mixed WTP results between on-site and self-administered surveys in a forest park in Scotland[23].

Finally, a few CV studies compare mail with telephone (Mannesto and Loomis 1991; Loomis and King 1994; Lindberg et al. 1997; Whittaker et al. 1998; Ethier et al. 2000; Kramer and Eisen-Hecht 2002)[24]. There is little testing and evidence of social desirability biasing WTP (see e.g. Legget et al. 2004 for a review of the older of theses studies), perhaps since such bias has traditionally been related to in-person interviews. Further, many of the studies acknowledge that sample composition effects cannot be distinguished from mode effects, so they focus more on assessing response rates, data quality in terms of item non-response and other sample biases between modes. They generally have little to say about response quality and satisficing, though two point out that increased time to think in the mail survey (provided such time is actually used) may make mail surveys more suitable for CV questions than telephone (Mannesto and Loomis 1991; Lindberg et al. 1997)[25]. Finally, meta-analyses of the valuation literature typically find differences between survey modes, e.g. that high-response mail surveys tend to give lower WTP than low-response surveys (due to higher inclusion of less interested respondents) and both lower WTP than in-person interviews (see e.g. Lindhjem (2007)). Lindhjem (2007) speculates that in-person interviews may lead to higher WTP simply due to respondents' better understanding of the good (and hence appreciation of the

[23] Studies investigating interviewer effects in CV, rather than comparing in-person interviews with other modes, include Bateman and Mawby (2004) (impact of interviewer dress appearance), Loureiro and Lotade (2005) (social desirability related to interviewer origin), and Mannesto and Loomis (1991) (degree of interviewer experience).

[24] Loomis et al. (2006) compare a video format with telephone interview (including a mailed information booklet) and find similar WTP for both samples and no difference in reasons for refusing to pay.

[25] Whittington (1992) finds that if respondents are given more time to think in an in-person interview setting they will bid less. This is similar to what Davis (2004) finds for her focus group respondents: when they are given the chance and time to revise their bid, they generally reduce it.

proposed policy at hand). However, empirical mode results in the meta-analysis literature do not seem to be consistent – except for documenting that such effects exist.

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. In the next section we propose a few 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 (like some of the reviewed studies do), 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

First, we investigate two common indicators in survey research of satisficing applied to our WTP question:

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[26].

Further, we assess two indicators of social desirability bias:

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 hurdle to state a reason of (strong) protest in an interview situation, compared to a more "safe" response that the respondent cannot afford the good, that it has no value or similar reasons. However, as has been discussed, the effects of the social desirability channels indicated in Hypotheses 3 and 4 may be considerably dampened by the potentially induced response honesty resulting from interpersonal trust with an interviewer.

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

[26] A stronger version of this hypothesis, increased tendency to choose midpoints in rating scales due to lack of motivation from an interviewer, was hypothesised by Krosnick and Chang (2001) to be observed in their Internet sample in a comparison with a RDD telephone sample.

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) (Bateman et al. 2002). 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[27], 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.

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[28]. Non-rejection of a traditional null hypothesis (to find a significant difference) 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, while 10 percent has also been used e.g. in clinical trials (Rogers et al. 1993). 20-40 percent has been suggested by Kristofersson

[27] 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.

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

and Navrud (2007) for benefit transfer applications in environmental economics. 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 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[29]. 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

[29] Admittedly, economic theory has relatively little to say about some of the variables typically included in validity regression analysis for CV (e.g. gender, attitudes etc). However, a growing body of CV research demonstrates that a wide range of these variables is empirically important in explaining variation in WTP.

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 less formally the share of respondents increasing their bid going from a small biodiversity protection plan to a larger one, i.e. the degree of internal scope sensitivity. It would be difficult to judge whether a higher share of bid increases 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 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 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 small 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 together[30]. 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 (2/3 in Southern Norway[31]), and the improvements in the living conditions for main groups of species[32]. The biological information was provided by a team of leading biologists in Norway, and checked by foresters to ensure a balanced 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

[30] It is recommended to give the respondent notice at the start that there are two plans to be considered (so-called "advance disclosure") (Bateman et al. 2004). It was not mentioned to respondents that the 4.5 percent plan was recommended by biologists, only that different plans were being considered.

[31] Since the exact location of future reserves is not yet decided the existing reserves in Southern Norway were scaled up on the maps to the correct relative size (2.8 percent and 4.5 percent of productive forest area, respectively), to give the respondent an idea of space requirements.

[32] Complex information about the environmental good such as this has been shown to be easier to comprehend if presented in a tabular format, such as the one used here, rather than in normal text (Hoehn et al. 2004).

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[33]) arranged on a non-linear scale in a table, including "don't know" (at the end). The amounts where chosen on the basis of previous CV studies (e.g. Lindhjem (2007)). PC was chosen as response format over dichotomous choice (DC) to preserve sample efficiency and because it lends itself nicely to the drop-down menu format very familiar to Internet-users. 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 payment vehicle (an annual, earmarked tax to a forest protection fund) was preferred because it is relatively realistic and reduces people's scepticism that the money would not be spent on forest protection. 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[34].

Survey administration in the two modes

A randomly recruited panel of 35000 willing respondents, maintained by the professional survey firm TNS Gallup, was used for the survey[35]. To the extent possible in a field experiment like this[36], confounding effects not related to survey mode was sought controlled

[33] There was also an option to choose "more than 15000", in which case a box would pop up in the Internet survey where the exact amount could be specified or the interviewer would note down the amount.

[34] The survey instrument, including a link to the Internet version, is available from the authors on request.

[35] TNS Gallup uses no form of self recruitment (e.g. through links on websites or similar), 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 etc.

[36] 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 looses some degree of control over influencing factors.

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[37]. 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 revealed) would be conducted by in-person interview and those able 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[38]. 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 typical Internet survey. The Internet sample was then recruited from the panel using the same e-mail except that the survey weblink was included so willing respondents could enter directly and reply to the survey[39]. Since the panel contains background information about all members, the

[37] In-person interviews are often sampled in clusters to reduce time and travel costs. In our case, this was not necessary. Instead after the sample was drawn, respondents living in similar areas of Oslo were identified and given appointments at successive times for home visits.

[38] 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 only the case for around 5 percent of the sample.

[39] 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).

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[40]. For the most important questions, including the payment card, reply options were given on display cards with the same appearance as on the Internet. Maps, colour photographs and graphs were displayed from an interview folder in the same order as in the Internet survey to avoid well-known response order effects, which depend on whether response alternatives are heard or read[41]. 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[42]. We also asked interviewers whether they thought respondents had understood the WTP questions. Theses 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

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

[41] 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).

[42] The question was answered on interviewers' pocket computer and was phrased in the following way: "To what extent do you think that the respondent felt the situation made it difficult for him/her to say no to supporting the program?".

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 20-30 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 protection policies; and the proposed policies and WTP questions[43]. Sixth, the surveys were conducted during the same time period of one and a half month (October-November 2007) to ensure preference stability and consistency between modes. Finally, the same token incentive to reply that are given for an Internet survey was also credited the in-person respondents[44], and all respondents were interviewed individually, not in groups of household members or similar[45]. Overall, the experimental design ensures a good opportunity for isolating 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 the results of the mode comparison.

[43] The Internet survey included an automatic timer that started and stopped when the respondent had read the information and answered related questions. Unfortunately, this information was not available for the in-person interviews.

[44] Credit points that can be used to buy various consumer products.

[45] The interviewers can control that this is the case in the in-person interviews, but 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, perhaps giving a higher degree of control than in standard mail surveys.

Results and analysis

Samples and response rates

The response rates for the in-person and Internet surveys were 59.7 and 75.4 percent, respectively[46], which compares favourably with similar surveys (even from pre-recruited panel respondents)[47]. 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 2007. Demographic variables are compared statistically between the two modes for both types of samples. Between the gross samples there is 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) 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.

[46] 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 approriate 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.

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

Table 1 Comparison of socio-economics across samples

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			$\chi^2=8.4^{*}$			$\chi^2=5.8$
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 in indicated income categories. 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 has to a larger extent responded in the interview mode). From inspection of Table 1 it can be seen that a slightly higher share of men has replied to the in-person survey compared to the Internet mode[48]. Non-response among groups according to age, income and education seems very similar for the Internet an in-person modes, respectively. In other words, type of survey mode does not overall seem to have influenced

[48] One speculative reason for this may be that the majority of our interviewers were women. However, as will become apparent in a later section, neither interviewer nor respondent gender influence WTP significantly.

whether people respond or not – both gross and net 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[49].

Satisficing & social desirability

We start by reporting the results from the satisficing hypotheses (H1 & H2). As noted, we use the first WTP question (small protection plan) as basis for testing our main hypotheses. 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 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 closer 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 53) (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 footnote 53.

[49] 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. However, on inspection of Table 1 such further analysis was left out for sake of brevity and simplicity.

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, as stated in one of the follow-up questions. 25 percent of Internet respondents and 17 percent of interview respondents stated this. Although the difference is not significant, it is an indication that interviewers make it easier for respondents to focus their attention and answer a difficult question. 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 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). That respondents with low education find it difficult to answer questions is of course problematic for survey based research in general, and may be particularly so for the more complex CV surveys. 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[50]. 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

[50] 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).

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 asked in standard CV fashion to state up to two reasons from a list of 11 possible reasons (including “don’t know”), 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 (see e.g. Bateman et al. 2002). In our case, this interpretation includes all reasons other than “I can’t afford to pay anything” and “Current level of protection is good enough”, which indicate “true zero”. 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 a share of protest of 90.65 for the interview sample and 88.06 percent for the Internet sample, a difference that is not statistically significant (see row seven in Table 2). The level of protesting is also surprisingly close between modes. 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 too state with "a straight face". Classifying only the former types of responses as strict protest gave a share of 74.77 percent protest in the interview sample and 70.90 in the Internet sample (row eight in Table 2). Somewhat surprisingly, protesting is now about 4 percentage points higher in the interview sample. The reason for this result may be found in the greater degree of interpersonal trust, and resulting response honesty, that can develop in an interview situation – though it is hard to pin this down in our data. In any case, the formal 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 it. These results run contrary to those of Legget et al. (2003) who find, using several logit models, indications of socially desirable responding to the 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 of) 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.

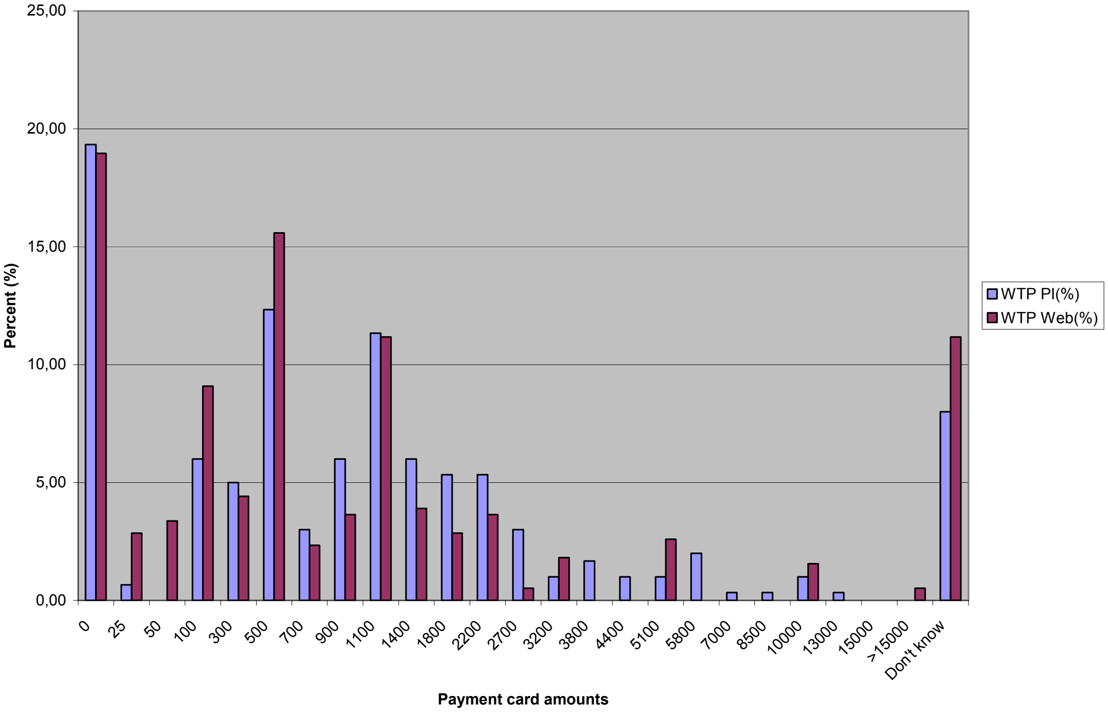
Overall then, little evidence has been found in our data for our hypotheses of social desirability bias and lower level of satisficing in the in-person interviews. The next step is to compare WTP between modes.

Comparison of mean WTP

The household 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) and Haab and McConnell (2002). Since the stated WTP amounts have a skewed distribution with the familiar long right tail, a log-transformation of WTP was applied[51].

[51] 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.

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

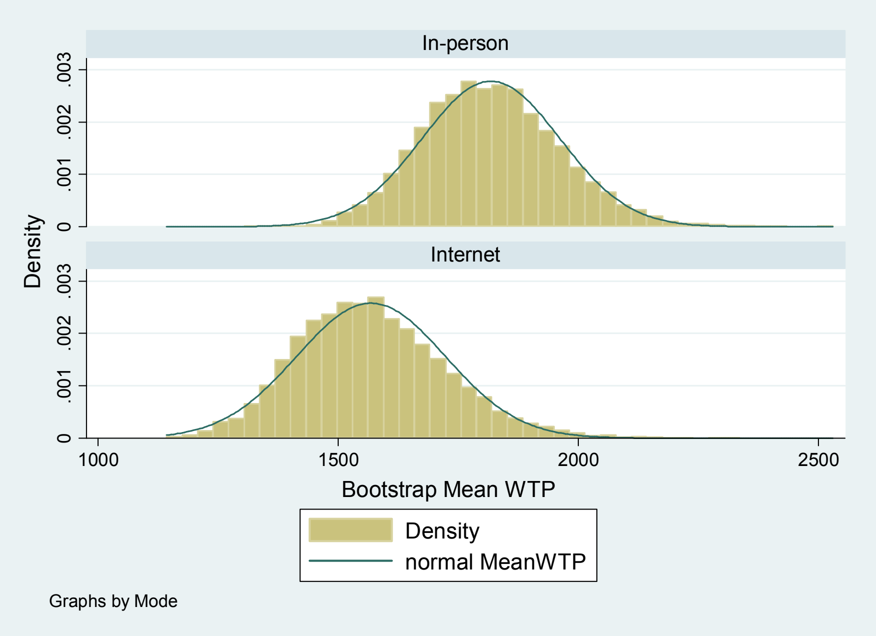
Table 3 Comparison of mean WTP between modes. WTP in Norwegian Kroner.

Hypothesis	**Interview: Mean WTP (95% CI) (n=218)**	**Internet: Mean WTP (95% CI) (n=269)**	**Mode comparison result (p<0.1)**
H5a Equality of means	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[52]. 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.

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

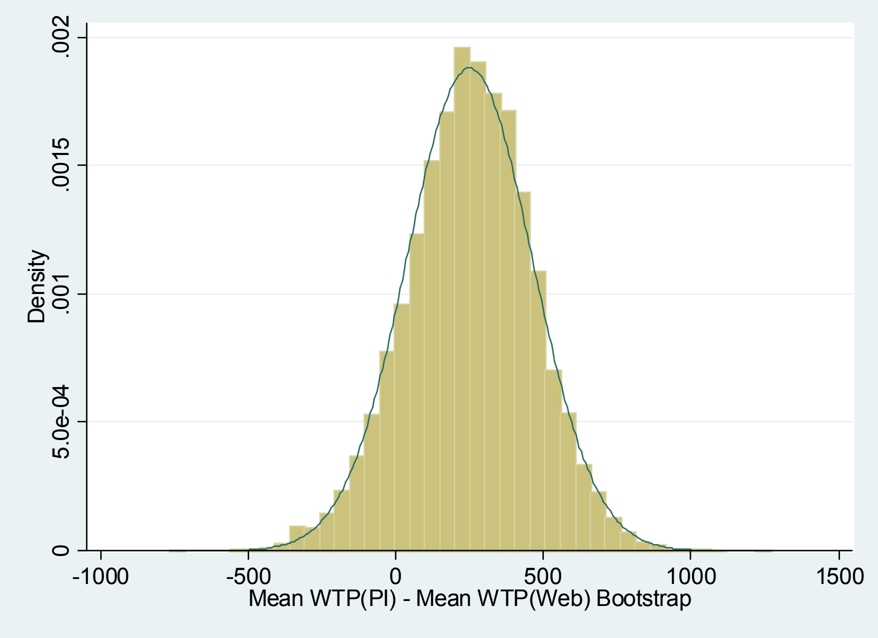


However, as we have argued, failing to reject the traditional null can in our case not be constructively interpreted as confirming convergent validity of WTP estimates between modes (in the same way a rejection cannot meaningfully be taken as evidence against convergent validity). Instead, we investigate whether the difference between means is of

[52] 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.

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 5 or 10 percent levels. 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. The same applies for a 10 percent equivalence bound (here 66 of the distribution is ≥ 10 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 hypothesis of 40 percent difference between means can be rejected at the 1.1 percent level (row seven). 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, i.e. the WTP difference is larger than 30 percent between modes. 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[53].

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 survey. c. A few respondents did not state income, so these observations have been excluded. Interval regression in STATA Version 9.2. used.

[53] Based on the results of a likelihood ratio test, we do not run pooled models. The likelihood ratio statistic is q=-2[logL$_{PooledAB}$ – (logL$_{A}$+LogL$_{B}$)~$\chi 2$ (d.f.), where logL$_{A}$ and logL$_{B}$ refer to the log likelihood values from the estimated models for WTP for individual samples (without covariates), and logL$_{PooledAB}$ 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 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

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 Legget et al. (2003) 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, if it were not for the fact that that time spent on the other parts of the survey has no effect on WTP[54]. Hence, it seems that if respondents can be motivated to spend time (and presumably effort) answering the WTP question, they will state higher WTP. This result is interesting and runs contrary to what was found by Whittington et al. (1992) and Davis (2004), where people typically revise their bid downwards when given more time to think (in an interview setting). Holgraves (2004) found that socially desirable responding was related to longer response times (not directly related to CV and WTP) (see footnote 16). 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 made a cursory check of whether people increase their WTP when an alternative, and larger protection plan of 4.5 percent of the forest is offered. While not performing a formal statistical test or internal scope test, 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. The internal scope validity seems to be similar between modes (and there seems to be no reason to suspect social desirability bias in people’s response to the second WTP question).

[54] Unless more interested respondents are also more knowledgeable and skim quickly through information they already know. However, judging from the type and detail of information provided, we think this is unlikely.

Concluding remarks

In a controlled CV field experiment of forest protection we have conducted the first test of whether results 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. Looking in particular for indications of social desirability bias and 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. Deciding the equivalence level is not straightforward. 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). Another, more explorative result is that the more time respondents spend answering the WTP question on the Internet, the higher is their stated WTP (while more time spent on other sections of the survey has no such effect). This result runs counter to some effects generally observed in interviews, where people typically revise their bids downwards if given more time (see e.g. Whittington et al. 1992).

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 (not least as documented by the research programs collected in Kahneman and Tversky (2000) and Gilovich et al. (2002)). 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 (as e.g. social desirability bias is 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). Relatively small samples are also a potential constraint. However, we are less concerned about the representativeness of respondents in our sample; they are unlikely to be very different from the general Norwegian population in terms of Internet use and familiarity.

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), Urama and Hodge (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 biased compared to in-person interviews. Further, if anything, our results show that the Internet mode gave slightly lower WTP. Since we don't know the true WTP of the respondents, it is important to estimate those values conservatively – in the spirit of the NOAA panel. 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 in future CV applications.

Acknowledgement

Funding from the Norwegian Research Council (grant 173230/I10) is greatly appreciated.

References

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

Andreoni, J. (1990), 'Impure Altruism And Donations To Public-Goods - A Theory Of Warm-Glow Giving', *Econ. J.* **100**(401): 464-477.

Arlinghaus, R. and T. Mehner (2003), 'Management and Ecological Note: Testing the reliability and construct validity of a simple and inexpensive procedure to measure the use value of recreational fishing', *Fisheries Management and Ecology* **11**: 61-64.

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

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

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

Bateman, I. J., D. Burgess, G. H. Hutchinson and D. I. Matthews (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., R. T. Carson, B. Day, W. M. Hanemann, N. Hanley, T. Hett, M. Jones-Lee, G. Loomes, S. Mourato, E. Ozdemiroglu, D. W. Pearce, R. Sugden and T. Swanson (2002), *Economic Valuation with Stated Preference Techniques: A Manual.* Edward Elgar Publishing, Cheltenham, 480pp.

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

Berrens, R. P., A. K. Bohara, H. Jenkins-Smith, C. Silva and D. L. Weimer (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., A. K. Bohara, H. C. Jenkins-Smith, C. L. Silva and D. L. Weimer (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.

Blamey, R. K., J. W. Bennett and M. D. Morrison (1999), 'Yea-Saying in Contingent Valuation Surveys', *Land Economics* **75**(1): 126-141.

Blomquist, G. C. and J. C. Whitehead (1998), 'Resource quality information and validity of willingness to pay in contingent valuation', *Resour. Energy Econ.* **20**(2): 179-196.

Boyle, K. J. (2003), 'Contingent valuation in practice', in P. A. Champ, K. J. Boyle and T. C. Brown, eds, *A primer on nonmarket valuation.* Kluwer Academic Publishers.

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

Cameron, T. A. and D. D. Huppert (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.

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.

Damschroder, L. J., P. A. Ubel, J. Riis and D. M. Smith (2007), 'An alternative approach for eliciting willingness-to-pay: A randomized Internet trial', *Judgment and Decision Making* **2**(2): 96-106.

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, eds, *Surveying subjective phenomena: Volume 2.* New York: Russel Sage.

Denscombe, M. (2006), 'Web-based questionnaires and the mode effect - An evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes', *Soc. Sci. Comput. Rev.* **24**(2): 246-254.

DeShazo, J. R. and G. Fermo (2002), 'Designing choice sets for stated preference methods: The effects of complexity on choice consistency', *J.Environ.Econ.Manage.* **44**(1): 123-143.

Dickie, M., S. Gerking and W. L. Goffe (2007). *Valuation of Non-Market Goods Using Computer-Assisted Surveys: A Comparison of Data Quality from Internet and RDD Sample.* Presented at European Association of Environmental and Resource Economists, Thessaloniki, Greece, June 2007

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

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

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

Edwards, S. F. and G. D. Anderson (1987), 'Overlooked biases in contingent valuation surveys: Some considerations', *Land Economics* **63**(2): 168-178.

Efron, B. (1997), 'Bootstrap methods: Another look at the jacknife', *Annals of Statistics* **7**: 1-26.

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

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

Fricker, S., M. Galesic, R. Tourangeau and T. Yan (2005), 'An experimental comparison of web and telephone surveys', *Public Opinion Quarterly* **69**(3): 370-392.

Gabaix, X., D. Laibson and G. Moloche (2003). *The allocation of attention: theory and evidence*, Working Paper, Harvard University

Gilovich, T., D. Griffin and D. Kahneman (2002), *Heuristics and biases : the psychology of intuitive judgement* Cambridge University Press, Cambridge.

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

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

Hanley, N. (1989), 'Valuing rural recreation benefits: An empirical comparison of two approaches', *Journal of Agricultural Economics* **40**(3): 361-74.

Harpman, D. A., M. P. Welsh and E. W. Sparling (2004), 'Unit non-response bias in the interval data model', *Land Economics* **80**(3): 448-462.

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

Hoehn, J. P., F. Lupi and M. D. Kaplowitz (2004). *Internet-based Stated Choice Experiments in Ecosystem Mitigation: Methods to Control Decision Heuristics and Biases*

Holbrook, A. L., M. C. Green and J. Krosnick (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.

Holtgraves, T. (2004), 'Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding ', *Personality and Social Psychology Bulletin* **30**(2): 161-172.

Hudson, D., L. Seah, D. Hite and T. C. Haab (2004), 'Telephone presurveys, self-selection, and non-response bias to mail and internet surveys in economic research', *Applied Economics Letters* **11**(237-240).

Haab, T. C. and K. E. McConnell (2002), *Valuing Environmental and Natural Resources: The econometrics of non-market valuation.* Edward Elgar.

Johnston, R. J. (2007), 'Choice experiments, site similarity and benefits transfer', *Environmental & Resource economics* **38**: 331-351.

Johnston, R. J., E. Y. Besedin, R. Iovanna, C. J. Miller, R. F. Wardwell and M. H. Ranson (2005), 'Systematic variation in willingness to pay for aquatic resource improvements and implications for benefit transfer: a meta-analysis', *Canadian Journal of Agricultural Economics* **53**(2-3): 221-248.

Jäckle, A., C. Roberts and P. Lynn (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,

Kahneman, D. and A. Tversky, Eds. (2000), *Choices, Values and Frames,* Cambridge University Press.

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

Kramer, R. A. and J. I. Eisen-Hecht (2002), 'Estimating the economic value of water quality protection in the Catawba River basin', *Water Resources Research* **38**(9).

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

Kristofersson, D. and S. Navrud (2007), 'Can Use and Non-Use Values be Transferred Across Countries?' in S. Navrud and R. Ready, eds, *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.

Krosnick, J. and L. Chang (2001). *A comparison of the random digit dialing telephone survey methodology with internet survey methodology as implemented by knowledge networks and Harris interactive*, Ohio State University

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

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

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

Lindberg, K., R. L. Johnson and R. P. Berrens (1997), 'Contingent valuation of rural tourism development with tests of scope and mode stability', *J. Agric. Resour. Econ.* **22**(1): 44-60.

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 S. Navrud (2008). *Asking for Individual or Household Willingness to Pay for Environmental Goods: Implication for aggregate welfare measures* Department of Economics and Resource Management, Norwegian University of Life Sciences

List, J. A., A. K. Bohara and J. Kerkvliet (2004), 'Examining the Role of Social Isolation on Stated Preferences', *American Economic Review* **94**(3): 741-752.

Loomis, J. and M. King (1994), 'Comparison Of Mail And Telephone-Mail Contingent Valuation Surveys', *J. Environ. Manage.* **41**(4): 309-324.

Loomis, J., J. Miller, A. Gonzales-Caban and P. A. Champ (2006), 'Testing the Convergent Validity of Videotape Survey Administration and Phone Interviews in Contingent Valuation', *Society and Natural Resources* **19**(4): 367-375.

Loureiro, M. L. and J. Lotade (2005), 'Interviewer Effects on the Valuation of Goods with Ethical and Environmental Attributes', *Environmental & Resource economics* **30**: 49-72.

MacMillan, D., N. Hanley and N. Lienhoop (2006), 'Contingent valuation: Environmental polling or preference engine?' *Ecological Economics* **60**(1): 299-307.

Maguire, K. B., G. Shiferaw and L. O. Taylor (2002). *Mode and subject pool effects in contingent valuation surveys using hurdle models. Working paper*

Malhotra, N. and J. Krosnick (2007), 'The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples', *Political Analysis*: 286-323.

Mannesto, G. and J. B. Loomis (1991), 'Evaluation Of Mail And In-Person Contingent Value Surveys - Results Of A Study Of Recreational Boaters', *J. Environ. Manage.* **32**(2): 177-190.

Marta-Pedroso, C., H. Freitas and T. Domingos (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., J. C. Bergstrom, C. M. Cornwell, R. J. Teasley and H. K. Cordell (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 R. T. Carson (1989), *Using Surveys to Value Public Goods: The Contingent Valuation Method.* Washington DC: Resources for the Future.

Navrud, S. and G. Pruckner (1997), 'Environmental Valuation - To Use or Not to Use? A Comparative Study of the United States and Europe', *Environmental and Resource Economics* **10**(1): 1-26.

OECD (2008). *OECD Broadband Portal.* http://www.oecd.org/sti/ict/broadband

Olsen, S. B. (2007). *Internet versus mail: Are stated preferences affected by the mode of sampling in a choice experiment?* Presented at European Association of Environmental and Resource Economists, Thessaloniki, Greece, June 2007.

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

Roger, J. L., K. I. Howard and J. T. Vessey (1993), 'Using significance tests to evaluate equivalence between two experimental groups', *Psychological Bulletin* **113**(3): 553-565.

Rosenberger, R. S. and J. B. Loomis (2000), 'Using meta-analysis for benefit transfer: In-sample convergent validity tests of an outdoor recreation database', *Water Resources Research* **36**(4): 1097-1107.

Sanders, D., H. D. Clarke, M. C. Stewart and P. Whiteley (2007), 'Does Mode Matter For Modeling Political Choice? Evidence From the 2005 British Election Study', *Political Analysis*: 257-285.

Schkade, D. and D. Kahneman (1998), 'Does living in California make people happy? A focusing illusion in judgments of life satisfaction', *Psychol. Sci.* **9**: 340-346.

Schulze, W., G. McClelland, D. Waldman and J. H. Lazo (1996), 'Sources of bias in contingent valuation', in D. J. Bjornstad and J. Kahn, eds, *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, eds, *The contingent valuation of environmental resources: Methodological issues and research needs.* Edward Elgar.

Schwappach, D. L. B. and T. J. Strasman (2006), '“Quick and dirty numbers”? The reliability of a stated-preference technique for the measurement of preferences for resource allocation', *Journal of Health Economics* **25**: 432-448.

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

Smith, V. K. (2000), 'JEEM and non-market valuation: 1974-1998', *J.Environ.Econ.Manage.* **39**(3): 351-374.

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

Statistics Norway (2008). *The Internet Poll, 4th Quarter 2007*

Thurston, H. W. (2006), 'Non-market valuation on the internet', in A. Alberini and J. Kahn, eds, *Handbook on contingent valuation.* Edward Elgar.

Tourangeau, R., M. P. Couper and F. Conrad (2004), 'Spacing, position, and order - Interpretive heuristics for visual features of survey questions', *Public Opin. Q.* **68**(3): 368-393.

Tourangeau, R., M. P. Couper and F. Conrad (2007), 'Color, labels, and interpretive heuristics for response scales', *Public Opinion Quarterly* **71**(1): 91-112.

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

Tsuge, T. and T. Washida (2003), 'Economic valuation of the Seto Inland Sea by using an Internet CV survey', *Marine Pollution Bulletin* **47**: 230-236.

Urama, K. C. and I. Hodge (2006), 'Participatory environmental education and willingness to pay for river basin management: Empirical evidence from Nigeria', *Land Economics* **82**(4): 542-561.

Welling, P. G., F. L. S. Tse and S. V. Dighe (1991), *Pharmaceutical Bioequivalence.* New York: Marcel Dekker.

Whittaker, D., J. J. Vaske, M. P. Donnelly and D. S. DeRuiter (1998), 'Mail versus telephone surveys: Potential biases in expenditure and willingness-to-pay', *Journal of Park and Recreation Administration* **16**(3): 15-30.

Whittington, D., V. K. Smith, A. Okorafor, A. Okore, J. L. Liu and A. McPhail (1992), 'Giving Respondents Time To Think In Contingent Valuation Studies - A Developing-Country Application', *J.Environ.Econ.Manage.* **22**(3): 205-225.