Using Internet in Stated Preference Surveys:

A review and comparison of survey modes

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

[a] Norwegian Institute for Nature Research (NINA), Gaustadalleen 21, N-0349 Oslo, Norway.

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

Forthcoming in the *International Review of Environmental and Resource Economics*

Date of draft: July, 2011

[1] Corresponding author: **henrik.lindhjem@nina.no**, Norwegian Institute for Nature Research (www.nina.no) and Vista (www.vista-analyse.no).

Abstract

Internet is quickly becoming the survey mode of choice for stated preference (SP) surveys in environmental economics. However, this choice is being made with relatively little consideration of its potential influence on survey results. This paper reviews the theory and emerging evidence of mode effects in the survey methodology and SP literatures, summarizes the findings, and points out implications for Internet SP practice and research. The SP studies that compare Internet with other modes do generally not find substantial difference. The majority of welfare estimates are equal; or somewhat lower for the Internet surveys. Further, there is no clear evidence of substantially lower quality or validity of Internet responses. However, the degree of experimental control is often low in comparative studies across survey modes, and they often confound measurement and sample composition effects. Internet offers a huge potential for experimentation and innovation in SP research, but when used to derive reliable welfare estimates for policy assessment, issues like representation and non-response bias for different Internet panels should receive more attention.

1   INTRODUCTION

One way the economics profession tries to support its self-proclaimed position as the only "hard" social science is by favouring new and sophisticated quantitative methods for recovering information from often poor data, over the less glamorous but essential groundwork of minimising and controlling survey errors in data collection. Economists valuing environmental goods using stated preference methods (contingent valuation (CV) or choice modelling (CM[2])) 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. This development is much due to the debate in the wake of the US National Oceanic and Atmospheric Administration (NOAA) panel report on the use of CV in natural resource damage assessments (Arrow et al., 1993). However, as the diminishing returns to yet another econometric method to analyse SP 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 data collection efforts to enable more robust insights from empirical analyses. This suggested shift is also underscored by the growing strength and relevance of behavioural economics to environmental benefit measurement and other areas of environmental economics (see List et al., 2004; Brown and Hagen, 2010).[3] Although current best practice SP studies generally are thorough in questionnaire development and testing, the choice of data collection mode – mail, face-to-face (f2f), telephone, Internet[4] or a mix – is typically made with comparatively little

---

[2] Often used as a catchall phrase for discrete choice methods that include among others choice experiments and conjoint analysis.

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

[4] Computers have long been used in survey data collection both in combination with face-to-face (f2f) interviews (so called CAPI – computer assisted personal interviewing) and telephone (CATI – computer assisted telephone interviewing). Our

consideration of its potential influence on sample composition and on how preferences are formed and stated.

A growing number of Internet-based SP studies of environmental goods (even high-budget 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. Tsuge and Washida, 2003; Berrens et al., 2004; Ladenburg and Olsen, 2008; Lindhjem and Navrud, 2009; Cai et al., 2010). While the mass exodus from traditional survey modes to the Internet in SP research is gathering pace, we think it is worth pausing to consider how this new mode may influence the derived stated preferences and welfare measures for environmental goods. The research on the effects of survey mode, and Internet in particular, on survey responses and data quality is attracting a great deal of attention in the survey methodology literature (e.g. Cooper, 2008; Couper and Miller, 2008; Baker et al., 2010). This fast-accumulating knowledge has yet to fully spill over into the SP literature, though the number of SP studies investigating survey mode differences is growing in the environmental and natural resource economics publication outlets.[5] The US EPA has recently also considered the implications of using Internet in SP surveys for environmental benefit measurement (Taylor et al., 2009).

We review the theory and evidence of mode effects in the survey methodology and SP literatures, summarize the findings, and point out implications for SP practice and research needs in order to better evaluate the increased use of Internet in SP surveys. The questions we attempt to answer in this paper include: Which mode differences can be expected from theory

---

main focus here is on self-administered surveys conducted on the Internet, usually while the respondent is in her home or workplace (interchangeably termed "Internet survey" or "Web survey" in this paper).

[5] A number of meta-analyses of the environmental valuation literature, not reviewed here, also document systematic, though not consistent, differences in welfare estimates depending on survey modes (e.g. Lindhjem, 2007; Barrio and Loureiro, 2010). Such studies establish correlations and are less well suited to investigate reasons for mode differences.

and empirical studies in the survey methodology literature? What are the experiences to date from the studies in the SP literature comparing Internet with other survey modes? To what extent do the studies avoid confounding sample composition effects from measurement effects? And finally: what are the implications for further SP practice and research?

Investigation of survey mode effects have become even more topical in light of the recent convergence in the SP literature towards the view that preferences are discovered or constructed by the respondent during the data collection process, rather than merely revealed or uncovered by it (e.g. Carlsson, 2010). This has been an uncontroversial point in psychology and survey methodology for a long time. Survey methodologists make the point that data is a product of the collection process, i.e. generated at the time of the interview or completion of the questionnaire, rather than just being "there" to be collected (implying that "data collection" is a misleading term) (Groves et al., 2004). More recently, environmental economists have come to view preferences as 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; MacMillan et al., 2006; Bateman et al., 2008). This "constructivist" viewpoint does not necessarily mean that there is no "true" value or no stable and coherent preferences to be measured, only that economists need to be more sensitive to the fact that "the construction process will be shaped by the interaction between the properties of the human information processing system and the properties of the decision task, leading to highly contingent decision behaviour" (Payne et al., 1999:245). In addition to the so-called measurement effects arising from the process of responding to questions in different survey modes, modes also result in important data differences related to sample composition (in terms of population coverage, sampling methods and non-response bias).

F2f interviews have been the recommended "gold standard" for surveys in general and SP research in particular (Mitchell and Carson, 1989; Arrow et al., 1993). Mail and to some extent telephone surveys have been much more used in practice; mostly for reasons of lower cost. The current trend in SP research, like in other survey based research, however, is to collect data using the Internet (Thurston, 2006). Sophisticated questionnaires can be delivered to large samples on record time at fairly low costs. Judging from the current growth in Internet penetration rates and use, Internet also 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)).[6]

This review is structured as follows. The next section first explains the main types of survey errors and reviews recent evidence from the broader survey methodology literature. Section 3 then reviews studies in the SP literature that have compared survey modes, with particular emphasis on Internet surveys. Finally, Section 4 concludes and synthesizes the implications for SP practice and further research.

2   SURVEY MODE EFFECTS – SOURCES AND TYPES

2.1   Sources of survey mode effects

Survey modes may give rise to different results since they, simply put (1) provide access to different types of people; (2) attract different types of respondents, and (3) elicit different

---

[6] According to Internet World Stats (2010), the EU has experienced a 258 percent growth in Internet usage over the last 10 years. Average Internet penetration is currently at 67.6 percent of the EU population, with the highest rates well-above 90 percent (e.g. in Sweden). All countries, except Romania, Portugal, Bulgaria and Greece, have penetration rates above 50 percent. For North America (excluding Mexico and the Caribbean) the corresponding figures are 146.3 percent growth in 10 years and penetration currently at 77.4 percent. 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 somewhat dated (much like similar concerns about telephone coverage 40-50 years ago).

responses (Jäckle et al., 2010). Figure 1 provides a standard overview of the main steps in the survey process (boxes in the figure) related to measurement of the construct(s) of interest in a sample of respondents drawn from a population. At each step errors may occur (ovals in the figure). The total survey error, that all SP surveys are subject to, can be grouped into classes of measurement and representation. Starting with measurement, the first step is the potential mismatch between the underlying construct and the way it is measured for each respondent (oval "validity"). The key construct in SP surveys is typically true willingness to pay (WTP) in CV or actual marginal trade-offs between alternatives (implicit prices) in common CM applications, though recreational use frequencies, attitudes, socio-economic data and other constructs (and their relationships with e.g. WTP) are often also of interest to a SP researcher (see box in the bottom right corner of Figure 1).

Figure 1   Survey steps and sources of errors from measurement and representation
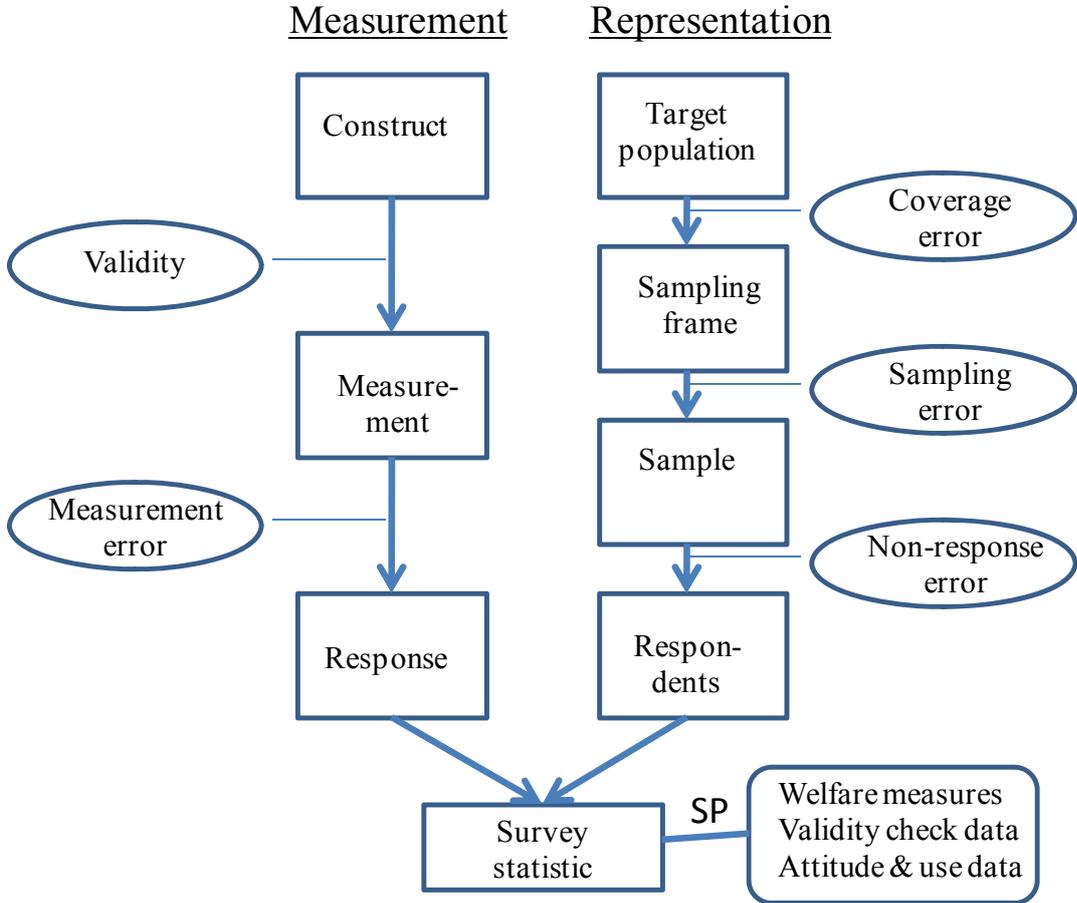
The actual or true benefit value is approximated through SP survey questions casted and answered in a hypothetical market setting. The *validity* of the measurement relies on the assumption that hypothetical answers relate closely to actual values or behavior. The validity of SP methods and potential hypothetical bias, especially in CV, has been discussed at length in the literature. The second source of error related to measurement is the mismatch between the ideal measurement of the sample unit and the response obtained – often termed *measurement error* (second oval to the left in the figure). The most important measurement error in relation to survey mode occurs when the same respondent provides different answers to survey questions that are worded the same across survey modes. This is sometimes regarded as the "pure" survey mode effect (Jäckle et al., 2010), and is of primary concern in this review.

The second source of potential survey errors related to representation come from the sampling of a limited number of respondents from a larger population (the right hand side of Figure 1). The three types of representation errors are *coverage error*, *sampling error* and *non-response error*. If the sample frame from which the sample is drawn does not match the population of interest one to one, this step introduces *coverage error* (first oval to the right in Figure 1). When drawing a sample from the sample frame all units may not have the same non-zero probability to be selected introducing *sampling error*. Finally, given that no surveys achieve response rates of 100 percent, the last source of error is related to systematic differences between actual responses of respondents and unobtained responses of non-respondents of

relevance to the constructs of interest; i.e. *non-response error*.[7] This error is related to self-selection bias, i.e. that respondents particularly (un)interested in a survey topic choose (not) to answer the survey.

The two processes of measurement and representation produce the survey statistics of interest for the relevant population, the basis in SP research for deriving welfare measures, investigate the construct validity of responses, and map attitudes and recreational use related to the environmental good (lower right box in the figure). Errors from measurement and representation are closely related. It is for example commonly assumed, though rarely tested, that people likely to become non-respondents to a survey are also likely to make lower quality responses if they do take part (Tourangeau et al., 2010). Some surveys use different modes ("mixed mode surveys"[8]) at different stages of the survey process, e.g. phone in recruiting respondents (or specific groups; e.g. the elderly) before subjecting all to an Internet survey. Our main emphasis in the following is on the potential effects of the choice of primary survey mode for administering a SP questionnaire. There are also potentially important differences within each mode. These include whether a f2f interview is conducted with the aid of a computer (CAPI), whether questions are read by the interviewer or by the respondent, whether the interview is conducted in-house or at a specific/centralized location like a recreational site, a shopping mall, a computer lab etc. It is also important that these aspects are kept in mind when analysing potential effects of survey modes, but we will try to distil some generic differences between telephone, Internet, mail and f2f interviews. In the next section

---

[7] Non-response is either the failure to complete a full questionnaire ("unit non-response") or to leave parts of the questionnaire unanswered ("item non-response"), in which case the questionnaire may have to be discarded if key answers are blank.

[8] See de Leeuw (2005) and Atkeson et al. (2011).

we start by discussing measurement effects, moving to errors related to representation in Section 2.3.

2.2 Measurement effects

Survey 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 normative factor is related to how cultural norms are invoked differently across modes leading to culturally constrained responses. The main difference is between a self-administered situation and the involvement of an interviewer. In addition, there may be smaller differences between mail and Internet on the one hand and telephone and f2f interviews on the other. The most important and well-documented mode effect in this regard, is according to Groves et al. (2004), the social desirability bias (DeMaio, 1984). This tendency of respondents to give the answer they feel they ought to give is sometimes more generally termed "compliance bias", and is more prevalent when an interviewer is involved (Green and Tunstall, 1999). The psychological factor is related to individuals' cognitive processing of information and questions, in particular how aural and/or visual stimuli produce different responses across modes.

When comparing measurement effects between survey modes due to the two main human factors, the consequences of satisficing (i.e. shortcutting the response process) and socially desirable responding are seen as central in the survey literature. Emerging research has

investigated different data quality indicators, for example completeness (e.g. item non response), accuracy (comparison with external benchmark data, e.g. on actual votes), reliability (e.g. psychometric scale properties) and more generally comparing response distributions of key constructs under study (Jäckle et al., 2010). Next, we discuss the degree of social desirability bias and satisficing for different modes, and review relevant empirical research.

2.2.1 *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. Respondents show this bias when they over-report socially approved behaviours or attitudes and underreport socially disapproved behaviours and attitudes. The response may be retrieved and then deliberatively edited after exerting much effort (Holtgraves, 2004), or be a result of shortcutting the response process and merely echoing what is thought socially desirable or politically correct. The extent of such responses seems to closely relate to two main factors: i) the degree of anonymity or "social distance", and ii) the trust or intimacy felt by the respondent while answering the survey. Social distance is minimised in a f2f interview 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 the highest for in-home interviews. Even if the respondent is allowed to submit a response anonymously (e.g. on a note put in a "ballot box", as suggested 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 topic at hand (Schkade and Kahneman, 1998) or get a slightly troubled conscience.

The cost of an honest, but socially undesirable response is the lowest when answering mail and Internet surveys, while telephone occupies a middle position. On the other hand, a great deal of interpersonal trust can emerge between an interviewer and the respondent in a f2f interview, especially in the respondent's home.[9] 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 and be more honest resulting in a lower number of socially desirable responses. 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 f2f surveys Holbrook et al. (2003) argue that the opposite effects of social distance and interpersonal trust on social desirability bias may cancel out in empirical applications.

Contrary to common beliefs, and those held by the NOAA panel (Arrow et al 1993), social desirability bias has often been found to be larger in telephone than in f2f interviews, at least for questions with some degree of sensitivity (see e.g. Groves et al. (2004) p158 or Jäckle et al. (2010)).[10] In addition to social distance and trust 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, e.g. attitudes towards the survey sponsor or topic (Tourangeau et al., 2009). In five recent papers, Internet was found to give lower degree of socially desirable responding compared to telephone interviewing (Kreuter et al., 2008; Chang

---

[9] F2f 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 nearby. However, little is known about effects of different types of f2f interviews or locations.

[10] 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)). Such effects have also been documented in the SP literature, see Section 3.3.

and Krosnick, 2009; 2010; Holbrook and Krosnick, 2010) and f2f interviews (Heerwegh, 2009).

The relative importance of the different effects related to social desirability in different modes discussed above is hard to assess for SP surveys. First, it is clear that since a SP 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 recreation user days over-reported in telephone or f2f interviews. Such biases may have implications for general assessments of the desirability of a proposed policy and for judging the validity of the SP data.

For CV the actual WTP question(s) can be influenced by social desirability bias since it may be considered a "civic virtue" (much like voting) to contribute to a common good. The effect may depend on the payment format (e.g. open ended, payment card – PC vs. dichotomous choice – DC). DC is likely to be more susceptible to yea saying, a well-documented problem (Blamey et al., 1999), in f2f or telephone interviews than in Internet or mail modes.[11] 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 stating higher WTP may be the most likely response. For both WTP formats it is unclear *a priori* how social desirability may influence incentive compatibility and strategic bias.[12] The degree of stated zero WTP and number of protest zeros

---

[11] 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.

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

can be expected to be lower if social desirability effects are at work. This is of direct importance to the estimation of mean WTP.

CM surveys are generally seen to be less susceptible to hypothetical bias than CV. However, as for DC they could be subject to affirmative behaviour, and it is not immediately obvious what is the socially desirable choice since the decision process is more complex than DC (and this complexity increases with the number of attributes and number of levels of each attribute).

Other standard SP questions such as the degree to which the respondent has understood the valuation scenarios or choice setting and whether he thinks the policy proposals are realistic – important for validity judgements of the data – may also not go free from bias. Finally, most of the background information collected in SP surveys will be truthfully reported regardless of mode (i.e. gender, age etc.), though some are typically not (especially income[13] and education). Based on expected mode effects discussed above, different measures of social desirability for the whole or parts of a survey (e.g. as an index[14]) or single questions could be constructed and tested. A few studies have investigated social desirability bias in CV, as reviewed in Section 3.

2.2.2 Satisficing

To execute the response process well, respondents need to exert some degree of effort and in SP generally more so than in other surveys (Mitchell and Carson 1989). CM surveys may as mentioned potentially reduce hypothetical bias problems prevalent in CV, but may also be regarded as even more challenging than CV surveys to answer, at least when the number of attributes and levels increase. Failure to put in the necessary effort to optimally answer a

---

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

[14] As suggested by e.g. Stocke and Hunkler (2007).

survey question, i.e. shortcutting the response process, leads to a satisfactory answer instead, or "satisficing". This term, originally coined in economics by Herbert Simon (1956), was first used in the survey literature 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.[15] Ability is often proxied fairly accurately by education level. People with less cognitive sophistication seem to be more affected by contextual cues when answering questions that are difficult to process (Toepoel et al., 2009). Ability is in turn closely related to motivation. When answering survey questions 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 SP surveys (and in other choice contexts) and allocate mental effort resources to this task, even though SP researchers for some time have studied framing effects, range and anchoring biases in the WTP response formats, and impacts on WTP (and other response variables) of varying the quality and quantity of information (and various stimuli such as e.g. colour photographs and video) (Navrud, 1997; Blomquist and Whitehead, 1998; Mathews et al., 2006). Promising research explicitly studying complexity, information processing and effort allocation include Berrens et al. (2004) and Lienhoop and Fischer (2009) for CV, Meyerhoff and Liebe (2009) and Boxall et al. (2009) for CM, and DeShazo and Fermo (2002) comparing both SP approaches. Gabaix et al. (2003) provides a more general theoretical framework.

---

[15] Although it is mostly assumed that satisficing increases monotonically with task difficulty, Malhotra (2009) argues that this view is too simplistic since respondents may be more motivated to complete tasks when they are intricate, challenging and enriching.

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. Schwappach and Strasman (2006) investigate a few such effects in their study of response reliability for an Internet CM survey. Measurement errors due to satisficing are sometimes difficult to separate from socially desirable responding[16] and response order effects not related to satisficing (Groves et al. 2004).

The main point here is how modes affect the tendency to satisfice for different types of questions in SP surveys. 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 interviews is the motivational effect of the interviewer. Green and Tunstall (1999) argue that in addition to practice (which is ruled out in most, standard "one-shot" SP 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.[17] These two factors reduce respondent benefits of

---

[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] Answers to questions respondents may have in SP 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

satisficing in interviews compared to the Internet mode. On the other hand a f2f 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. They 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 – or even virtual reality visualisations[18] – can be provided more easily etc., and the respondent can answer in her own time. The net effect for Internet and interview modes may be difficult to assess for SP surveys, although it is generally agreed that satisficing may be a bigger problem in self-administered than interview surveys (Holbrook et al., 2003). A few recent studies have assessed degree of satisficing and related data quality aspects for web surveys compared to other modes in political and other social science research. Chang and Krosnick (2009; 2010) found less satisficing in the Internet mode compared to telephone interviews, though some of the results were sensitive to the kind of Internet panel respondents were drawn from (see Section 2.3). Fricker et al. (2005) find mixed results in their comparison with a telephone survey. Heervegh (2009) and Heervegh and Loosveldt (2008), on the other hand, find indications of a higher degree of satisficing (e.g. more "don't know" and less differentiation on rating scales) in a web survey compared to f2f interviews. Malhotra and Krosnick (2007) find in a survey comparison of voter attitudes and behaviour in the U.S., lower accuracy in the data from the Internet volunteer panel than in the f2f survey. In contrast to this result, Sanders et al. (2007) found

---

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)).

[18] See Bateman et al. (2009).

few statistically significant differences between coefficients generated using f2f and Internet data.

Without considering satisficing explicitly Borkan (2010) found no mode effects between mail and Internet surveys on psychometric quality of rating scales and data quality (measured as item non-response). Also comparing with mail, Denscombe (2006) supports this finding and states that "there is little evidence of a mode effect linked to web-based questionnaires". Denscombe (2009) finds almost the same item non-response rates to fixed-choice questions, and lower item-non response to open questions in the web survey. In contrast with the above research, Rookey et al. (2008) found that web and mail respondents provided different answers to almost one third of the survey questions, with notable differences in opinion and behaviour questions. However, the degree of experimental control in terms of ability to disentangle measurement and sample composition effects varies in this literature. Apart from these studies, measurement errors between modes due to satisficing or for other reasons[19] have generally not been much studied for Internet surveys (Dillman and Smyth 2007). However, impacts of questionnaire design elements have been shown to be similar to mail surveys (Tourangeau et al. (2004; 2007), Galesic et al. (2008), Cooper (2008) and Dillman et al. (2009)).[20] Based on this limited research, it is not possible to conclude that satisficing generally is a bigger problem, or data quality substantially lower, for Internet surveys compared to other modes.

---

[19] For example, the way the aural vs. visual senses are stimulated may result in different processes through which the meaning of a question and the response alternatives are comprehended.

[20] 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).

Similar to the discussion for social desirability bias, different types of SP questions will be susceptible to satisficing in different ways, with the WTP or choice set questions obvious victims. For payment card in CV, satisficing may for example 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 (though actual protesting may be influenced by social desirability effects).[21] For CM surveys, indications of satisficing may for example include ignoring attributes (Carlsson et al., 2010) and various types of choice inconsistencies (DeShazo and Fermo, 2002).

2.3 Sample composition effects

Observed differences between modes in non-experimental studies may also be due to sample composition effects rather than differences in measurement per se, as depicted in Figure 1. In other words, observed differences may be due to *who* respond, rather than *how* they respond. For Internet surveys non-coverage (lack of Internet access or limited use) of the general population and high non-response (unwillingness to participate given access) are seen as the major challenges (Couper et al., 2007). A suitable sample frame for the general population (rather than a special-purpose population such as employees in an organisation) using e-mail does not exist for Internet surveys the way it does for other modes (Couper and Miller 2008).[22] Despite high and growing Internet coverage, certain groups, typically the elderly, people in rural areas and people with low education (and income), are currently underrepresented. Further, for Internet, as for other modes, the willingness to participate in surveys is declining creating potential non-response and self-selection biases that may vary

---

[21] A quick zero response may of course not necessarily imply satisficing, as many respondents may be very sure about such a response.

[22] The situation for other modes is, however, not static. The move towards increasing mobile phone use, for example, may generally make it harder both to sample people and to conduct the interviews in a suitable setting. Generally, people tend to get harder to contact for surveys, and when contacted, more reluctant to take part (Grandjean et al. 2009).

between survey modes (Groves, 2006). In addition, non-response seems to be more prevalent in Internet surveys (Manfreda et al., 2008; Shih and Fan, 2008). Non-response bias confounds mode comparisons when the (unobservable or observable) characteristics of people who prefer one mode to the other are correlated with the constructs researchers want to measure in the survey (e.g. WTP).

According to Couper and Miller (2008) two broad approaches have been developed to deal with the problems of coverage and non-response in Internet surveys. The first has been to attempt to build probability-based Internet panels of willing respondents by using other methods for recruitment and sampling (e.g. random digit dialling - RDD). An example of this approach in the USA is Knowledge Networks, which use RDD recruitment and provide free Internet (or web TV) access in exchange for joining the panel. The second approach involves recruiting willing respondents through different non-probability based means, for example through weblinks, advertisements etc. This kind of web-panel is termed "opt-in panel" or "volunteer panel". Even though samples from this type of panel may appear representative on socio-economic variables, statistical inference on such samples is unfounded as long as the initial selection is non-random. In order to reduce potential representational biases, quota sampling into subgroups (according to e.g. age, education, income levels etc) or post-survey weighting strategies are often used to better resemble the general population.[23] A review of empirical findings and recommendations for the responsible use of opt-in panels have recently been published by the American Association for Public Opinion Research (Baker et al., 2010). For both types of panels the real cumulative response rate (rather than just final-stage

---

[23] A criticism sometimes levelled against Internet panels of both types is that survey panellists may provide different responses than the average person, because panellists often respond to many surveys. Such effects are sometimes referred to time-in-sample or panel attrition effects.

for a survey administered to the panel) is often low and/or unknown and the identification of non-response biases is in its infancy (Couper and Miller, 2008).

One important advantage of Internet panels over other modes, offsetting some of the problem with likely lower response rates, is that they usually contain updated background information related to socio-economics, attitudes, political affiliation etc. for both respondents and non-respondents to a survey (though not about those that did not agree to participate in the panel in the first stage recruitment). This information can then be used to identify and correct for non-response bias according to observable characteristics (Heckman, 1979). However, if the non-response bias is uncorrelated with the observable background variables, such post-survey adjustments are not possible.

Few studies we are aware of in the survey methodology literature make stringent comparisons of sample composition effects between types of Internet panels and other survey modes. One exception is Yeager et al. (2009) who compare the accuracy of survey responses from two probability samples (Internet panel and RDD samples) and a non-probability recruited Internet panel. Benchmarks derived from official government records or high quality federal surveys were used to judge accuracy or validity of survey responses. They found that the probability sample surveys were consistently more accurate and that the post-stratification weighting applied to the non-probability sample improved the accuracy of some measures and decreased the accuracy of others. Hence, weighting may not be a reliable strategy. For SP surveys, the external validity is of course hard to judge as real WTP or the actual choice is not available. Both types of Internet panels have been used in mode comparisons in SP research, as have strategies of general e-mailing. In the next section we present and discuss experiences from these SP studies.

3   SURVEY MODE COMPARISONS IN THE STATED PREFERENCE LITERATURE

3.1   The survey mode debate in CV

In their landmark book on CV Mitchell and Carson (1989) argued that the mode of choice for CV surveys is f2f interviews conducted in the respondent's home, for three main reasons:

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.

The former two points are related to alleviating satisficing through making a complex task simpler and through motivation. The third point stresses the need to alleviate non-response. Mitchell and Carson (1989) do 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 Mitchell and Carson's main 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).[24] The NOAA panel, however, recommended controlling for interviewer effects, especially social desirability bias. Schuman (1996) (the survey expert on the NOAA panel) defended and explained the NOAA panel recommendation of f2f interviews.

---

[24] 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 SP research more generally (see e.g. Navrud and Pruckner (1997)).

Mail survey proponents such as Don Dillman strongly disagreed (see letter annexed in Schulze et al. (1996)). Schulze et al. (1996) called for more research comparing effects of different modes before definite recommendations for CV can be made. With the introduction of Internet surveys and the general increase in the use of CM methods, this call for more research seems still valid for SP methods. What is known about measurement and sample composition effects in Internet SP surveys compared to other modes?

3.2 Studies comparing Internet with other modes

In the following we first review the emerging evidence from Internet survey mode comparisons in the SP literature. We emphasise whether studies aim to investigate comparability of data in different modes and/or test hypotheses about potential causes of such effects (and especially whether sample and measurement effects are confounded), which survey questions or validity issues are compared and how, whether sample differences are discussed and (if possible) corrected for, and which conclusions the authors arrive at. Second, we briefly review other mode comparison studies in the SP literature that do not use Internet, but may still yield lessons of relevance for the analysis of Internet surveys.

3.2.1 Overview of studies

Table 1 provides the summary details from 17 identified SP studies that have compared Internet with other survey modes for environmental goods or environment-related health risks. 12 of these are CV studies, reflecting the still early days for CM research in environmental economics.[25]

---

[25] Including revealed preference methods, we have also identified two travel cost studies comparing Internet samples with either on-site (Hynes and Hanley, 2006) or mail surveys (Fleming and Bowden, 2009). Both these studies find similar responses and welfare measures between modes.

Table 1   Internet survey mode comparisons in the SP literature

Reference	Comparison with mean web WTP	Method[£]	Good valued	Key study issues
Banzhaf et al. (2006)[¤]	= mail	CV: DC	Ecological improvements	Weighted samples, check of non-response, panel attrition
Bell et al. (2011)	>central location <mall<phone-mail	CM[*]	Water quality	All computer administered. Focus on recruitment mode & sample comp.
Berrens et al. (2003)	None conducted	CV: DC	US Kyoto Protocol ratif.	Phone. Weighting, several issues compared; low response rate
Canavari et al. (2005)	> f2f (in a store)	CV: OE/DC	Organic fruit	Low response rate, self-selection acknowledged
Covey et al. (2010)	= f2f (in-home)[§]	CM[§]	Rail safety	Rating, ranking and matching questions studied
Dickie et al. (2007)	None conducted	CV: DC	Skin cancer risk	Computer at central location. Very different samples.
Grandjean et al. (2009)[§#]	= mail < phone	CV: DC	Clean air in national parks	Measurement effects; social desirability bias; weighting
Hudson et al. (2004)	None conducted	CV: OE	Water quality	Mail. Self-selection and non-response issues investigated
Li et al. (2009)[¤]	= phone	CV: DC	Energy R&D	Median WTP. Mode as dummy in Bayesian & standard WTP modeling
Li et al. (2004)[**]	= phone	CV: DC	US Kyoto Protocol ratif.	Equivalency of underlying preferences tested
Lindhjem & Navrud (2011)[§]	≤ f2f (in-home)[@]	CV: PC	Forest biodiv.	Measurement effects; satisficing & social desirability bias.
MacDonald et al. (2010)[§]	< mail	CM	Water quality	Both measurement and sample composition effects
Marta-Pedroso et al. (2007)	< f2f (on a beach)	CV: OE	Landscapes	Low web response rate. Very different samples
Nielsen (2011)	= f2f (in-home)	CV: OE	Clean air	Sample & measurement issues Surveys two years apart.
Olsen (2009)	= mail	CM	Landscapes	Sample & measurement issues. Protesting, respondent certainty
van der Heide et al. (2008)	= f2f (on-site)	CV: DC	Habitats	Construct validity checked, self-selection
Windle & Rolfe (2011)	= paper-based	CM	Coral reefs	A drop-off/pick up method. Attitudes, use, protesting.

Notes:

£ DC = Dichotomous choice; OE = open ended; PC = Payment card; CAPI = Computer assisted personal interview.

¤ Study that reports some mode comparison results, but did not have that as the primary purpose.

*Iterative choice between living in two regions that only vary in living costs and water quality.

$ Person trade-off or matching technique to estimate valuation rations (rather than WTP) used in the risk/value of statistical life literature.

§ Higher degree of experimental control than the other studies in terms of distinguishing measurement from sample effects.

Working paper based on the US EPA-funded study by Taylor et al. (2009).

** The same samples as in Berrens et al. (2003; 2004).

@ The authors reject that mean WTP for f2f is larger than 30 percent.

Six of the studies compare Internet with f2f interviews in various locations, five with mail or

paper-based surveys, three with telephone interviewing, one both with telephone and mail,

one with mall-intercept, phone-mail and central location samples, and finally one with a centrally-administered computer survey. The goods valued vary widely and cover water and air quality and environment-related health risks, nature protection and landscape amenities, and climate and energy policies (column four). The most common elicitation format used in the surveys is dichotomous choice (DC) CV (column three).

The second column of Table 1 shows the results of the 17 pair wise comparisons of mean WTP between modes for the 14 studies that have done this explicitly. Ten comparisons fail to reject any difference between Internet and the other modes. Of the remaining seven comparisons five give lower WTP for the Internet mode and two higher. Hence, the large majority of the studies find equal or lower welfare measures for the Internet mode. A closer look at each study is required to say anything about trends related to types of survey modes. The studies are generally fairly heterogeneous in design and in the choice of methods and issues compared across modes (see column five). It is therefore difficult to generalise from these studies. As we discuss next, the degree of experimental control also vary between the studies and it is often hard to tell whether observed effects (or lack thereof) are due to measurement or sample composition effects.

3.2.2 *Face-to-face comparisons*

Six studies to date have compared Internet with f2f interviews and Canavari et al. (2005) is to our knowledge the first. The study investigates Italians' WTP for a ban on pesticides in fruit production sampling customers of four large retail outlets in Bologna for the f2f interviews (conducted in the stores) and e-mailing a sample of the members of a community e-mail network in the same city for the web survey (the response rate for this survey was 6 percent). The two samples turn out to be very different in terms of common socio-economic characteristics. The Internet sample has high income, education and male overrepresentation

25

reflecting the unequal adoption of Internet in Italy. This factor, in addition to potential self-selection caused by the different sampling strategies, is the likely reason for the observed higher mean WTP in the Internet sample. The authors find, interestingly, that WTP from both samples vary in the same expected way to relevant socio-economic covariates.

Marta-Pedroso et al. (2007) sample visitors to a beach for interviews (conducted by the authors) and Internet respondents recruited via an e-mail list. They find about the same share of zero WTP and protests for the two modes for an environmental preservation program in Portugal. However, the mean WTP was found to be higher for the f2f 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 f2f mode is an indication of social desirability bias, probably made worse since professional interviewers were not used. However (similar to Canavari et al.), there are too 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 satisficing or other measurement issues in the study. As such it is more a practical comparison of modes than a controlled experiment.

Instead of using general e-mailing, Van der Heide et al. (2008) draw a sample from an Internet panel[26] of the Dutch population to compare with an on-site interview sample from the Veluwe region. The survey valued two scenarios to remediate habitat defragmentation in the region. The authors find that both samples are quite representative of the Dutch population. The survey results show almost identical rates (ca 27 percent) of "no-no" responses to the two DC valuation scenarios between the two samples. Hence, there seem to be no obvious signs of social desirability bias. For the WTP comparisons, they conclude that (p 213): "[We cannot]

---

[26] It is unclear whether it is a probability-based or an "opt-in" panel (see Section 2.3).

indisputably reject the hypothesis that WTP values derived through interviews are the same as values obtained from the Internet survey." In terms of judging construct validity, they find that WTP vary in expected and very similar ways for both samples. Hence, there is no indication of lower quality data due to satisficing in the Internet sample. The first point to note is that the two populations are different, as the Internet sample contains both users and non-users and the on-site sample users only. As it may be likely that users also hold higher non-use values than do non-users, one would expect higher WTP for the on-site sample. This issue is not analyzed in the paper. Another issue, alluded to in the conclusions, is that the on-site sample may have been more prone to self-selection than the Internet sample.

Nielsen (2011) aims to investigate both sample composition and measurement effects in a CV survey of life expectancy gains from air pollution reduction in Copenhagen, Denmark. Although the study has problems distinguishing the two effects from each other in practice, the paper adds some more conceptual clarity on survey mode differences than the other three papers reviewed above. The Internet sample was recruited from a probability-based Internet panel (final stage response rate of 40 percent), and the f2f sample recruited a year later and conducted in respondents' homes (70 percent response rate[27]). In terms of gender, age and income, the samples are fairly similar, though there are more highly educated people in the Internet sample (as noted in Section 2.3 is often found to be the case). The study further finds significantly more protest votes in the web sample, though real zero WTP responses are almost identical. Less protesting could be a social desirability effect in the f2f mode. To weigh up for the educational difference, a smaller sample with the same educational profile was drawn randomly from the Internet sample. The mean and median WTP from both these samples were found to be statistically indistinguishable from the f2f sample. Further, both

---

[27] It is unclear if this includes all people invited for the interviews, which would be the true response rate, or just the share of those who said yes and then did not end up actually taking the interview (most likely the latter).

samples satisfy the scope test and show similar validity in regressions, showing no clear indications of lower quality Internet responses.

Covey et al. (2010) report the results from a high-budget SP study estimating the valuation ratios for a range of different types of rail victims and accident situations using an approach known as "person trade-offs" or "matching" (classified here as a CM approach). The study was designed specifically to compare modes and therefore controlling confounding factors to a larger extent than the other studies above. The study used the same questionnaire, administered around the same time period and used professional survey firms for sampling and interviewing. The two large (>1000 respondents) f2f and Internet panel samples were comparable to each other and to the UK population they were drawn from (except for slightly younger respondents in the Internet sample). The first part of the survey contained rating tasks that yielded consistent results between the samples. The authors do, however, note that there may have been a greater tendency for the Internet respondents to rate different programs "equally good", i.e. interpreted as an unwillingness to make trade-offs. This may be an indication of satisficing or "lack of effort on the part of the respondents" as suggested by the authors. However, such results may also reflect genuine indifference, a position perhaps harder to put forward in an f2f setting. Covey et al. conclude that (p 85): "there was an encouragingly close correspondence between the findings of the face-to-face and internet surveys".

Finally, in a more classic CV setting Lindhjem and Navrud (2011) compare an Internet panel and f2f sample asked to value a high non-use value good: establishment of forest reserves to protect biodiversity in Norway. The study aims in particular to investigate measurement effects from both satisficing and social desirability response behavior. Sample composition effects are sought better controlled than the other studies through drawing the samples from

the same sample frame, namely the Internet panel of respondents. Respondents are invited by e-mail to either a f2f or Internet mode, so there may have been some self-selection. Still, this procedure generates very similar samples in terms of gender and age, though there are some small differences in education and income distributions. Not weighing for these differences, results of the mode comparison show little evidence of social desirability bias in the f2f setting or satisficing in the Internet survey. The share of "don't knows", zeros and protest responses to the WTP question with a payment card was found to be very similar between modes and equality of mean WTP between samples could not be rejected. Validity of the responses in terms of relationships of covariates with WTP was also shown to be similar for the two samples.

3.2.3   Mail, telephone and computer at central location

The remaining 11 studies include five comparisons with mail, three with telephone, one with both mail and telephone, and two including computers at a central location in combination with different recruitment strategies.

Reviewing the mail comparisons first, 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 were Internet panellists (from Knowledge Networks) and one third was mail respondents (from two different sample 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. They also sent their questionnaire to former members of the Internet panel to test for panel attrition effects, which is an interesting idea though results from this test are not reported.

Macdonald et al. (2010) aim to disentangle sample composition and measurement effects in a CM approach valuing river water quality in Australia. Like Grandjean et al. (2009) (reviewed

29

below) and Lindhjem and Navrud (2011), the paper attempts to heighten the experimental control in its mode comparison compared to previous studies. They do this by drawing two mail samples for comparison: one from Australia Post and one from the same sample frame as for the Internet panel sample. Other design features, such as incentives for answering the survey, questionnaire illustrations etc. were kept as similar as practically possible. The response rates were 57, 52 and 31 percent, for the two mail samples and the Internet sample, respectively. The Internet sample yielded somewhat younger, wealthier and more educated respondents compared to both mail samples. Environmental attitudes were found to be similar between samples. Estimating the implicit prices on water quality improvements, they find lower values for the Internet sample, indicating a survey mode effect. However, the study does not attempt to weigh for socio-economic differences that still prevail even though sample frames are the same.

Olsen (2009), also in a CM setting, investigates preferences for protecting recreational use values from motorway encroachment in two municipalities in Denmark comparing an Internet panel sample with a general mail sample. The survey achieves response rates above 60 percent for both samples. 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. The two samples differ predictably along dimensions of age, but have similar distributions for income and education. However, both samples have an overrepresentation of the high income, high education groups compared to the general population, a common feature in surveys. Comparing mean WTP between samples Olsen (2009) concludes that it cannot be rejected that preferences from the two modes are identical, though estimation precision and reliability of choices are higher in the mail sample. Finally, comparing reported respondent certainty, Olsen (2009) finds, interestingly, that this is significantly higher in the Internet sample. He

speculates that this may be due to survey experience of Internet panellists, though this is not formally investigated.

Following a similar approach to Olsen (2009), Windle and Rolfe (2011) compare an Internet panel sample with a drop-off/pick up paper-based collection method (details of which are unclear). The object of valuation is the improvement of the environmental condition of the Great Barrier Reef in Australia. Sample composition is similar in terms of income and education, though the Internet sample has younger and more male respondents. No sample was clearly more representative of the general population. The paper-based sample had more item-non response, both about recreational use and income data. Mean household WTP and level of protesting were found to be the same for the two samples. Despite this result, comparison of statistical modelling, attitude and use data, indicate some differences between samples, though in this study too, it is unclear if they are due to sample or measurement effects.

In a short paper, Hudson et al. (2004)[28] are primarily concerned with investigating self-selection and non-response bias in two mail and Internet samples. These main surveys are preceded by a telephone survey collecting socio-economic background information about potential respondents and responses to a simple CV WTP question, and asking for participation in a follow-up survey (by mail or Internet only, or left as a choice for the respondent). Their comparison of respondents vs. non-respondents from the telephone survey finds lower mean and median WTP (also when controlling for income) and lower income for the sample that refused to take the full survey, indicating selection-bias. A final check involved comparing respondents and non-respondents to the follow-up survey.[29] Despite

---

[28] Note that the paper confuses the terms "item- and unit-nonresponse bias" in the abstract.

[29] In other words, those who by phone had accepted to take part in the survey did not all complete the follow-up survey.

much lower response rates for the Internet sample, there was no clear indication that this sample was more prone to non-response bias, at least as judged by common socio-economic characteristics.

Grandjean et al. (2009) compare both mail and telephone samples with a probability-based Internet panel sample (Knowledge Networks), in the perhaps most comprehensive and well-controlled survey mode study in the SP literature to date. It was funded by US EPA recognising the need for more knowledge of whether Internet panels can produce reliable value estimates. In a fairly standard, national-level CV survey using a DC response format respondents were asked to value plans to reduce ground-level ozone concentrations in national parks. The three samples were drawn using RDD[30], ensuring a constant sampling frame. Efforts were also made to harmonise questionnaire design between modes as much as practically possible. A random within-household sampling method was chosen for the three modes (the adult person with the "last birthday" was chosen), as this may also be a potential source of differences in stated values (Lindhjem and Navrud, 2009). To further isolate measurement effects the samples were weighted to a set of marginal distributions on common demographics (though, importantly, not income), using census benchmarks. This was also done due to low response rates (as low as 4 percent for the Internet sample). Most likely since Internet respondents had already agreed to participate in the panel, less self-selection related to recreation behaviour and environmental attitudes was found for this mode. Compared to web panellists, phone respondents were more likely to accept bids, translating into 3-4 USD higher WTP on average than for the Internet respondents. The authors argue that this may be an indication of social desirability bias. There was no difference between mean WTP for mail and Internet respondents. The authors conclude that WTP estimates from a probability-based

---

[30] At the time of the study, the sampling fram for RDD covered 82 percent of all U.S. households.

Internet panel are no less accurate than for a well-executed mail survey, and probably more accurate than a phone survey.

There are three closely related studies comparing phone and Internet samples. Mainly addressing the issue of representativeness of two types of Internet samples[31] compared with a RDD telephone sample for political research, Berrens et al. (2003) also assess questions of environmental attitudes and WTP using a CV survey. 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). Utilizing the same dataset, Li et al. (2004) follow the approach suggested by Cameron et al. (2002) to determine if the underlying preferences from the Internet and telephone samples are the same. They conduct several tests, which support the hypothesis of equivalent WTP values and underlying preferences. Some of the same authors conducted another CV survey of WTP for increased energy research in the U.S. confirming the previous result of no difference in responses across two Internet panel and telephone samples (Li et al. 2009).

Finally, there are two studies that are slightly different in that they compare locations for computer surveys, in combination with different recruitment strategies. In a CV survey of reduced skin cancer risk Dickie et al. (2007) compare a sample recruited through a RDD procedure answering the survey on a computer in a central location with a sample of Internet panellists, collected three years later, and answering on-line. Their results suggest lower quality of responses for the Internet survey, indicating greater satisficing (though the authors

---

[31] One sample comes from Harris Interactive (an opt-in panel) and one from Knowledge Networks (a probability-based panel).

33

do not use this term). Internet respondents had more item non-response, rushed through the survey, 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 time-in-sample effects. 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).

Finally, in a recent study Bell et al. (2011) compare results from a CM type approach that make respondents choose iteratively between different combinations of water quality levels and living costs in a hypothetical move to another region. Four different modes of recruitment to a computer-based survey are compared: phone-recruitment to complete survey at a central location; mall intercepts to complete survey on computers in the mall; national phone-mail survey where recruitment was done my phone and a disc mailed to respondents for completion on own computers; and an Internet panel by Knowledge Networks. The study finds fairly large differences in valuation between recruitment modes, persisting also after demographic and other differences are controlled for in model estimations. There is, however, a number of confounding factors in the study including different survey years and questionnaires, making it difficult to draw firm conclusions.

3.3   Other mode comparisons in the SP literature

The involvement of the interviewer is as previously noted a potentially important source of measurement differences between self-administered Internet and mail surveys on the one hand and f2f and telephone surveys on the other. Our review above demonstrates a somewhat greater correspondence between mail and Internet survey results than Internet and f2f/telephone results, especially if the high-quality study of Grandjean et al. (2009) is given extra weight. Avoiding social desirability bias while keeping satisficing effects at acceptable levels is an important consideration in the choice of an Internet survey over interviews. The SP literature prior to the onset of Internet surveys has shed some light on the extent of measurement effects, especially social desirability bias, between traditional survey modes. We review some of these studies below.

List et al. (2004) find in a f2f 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 anonymity is increased. It is uncertain how this result can be transferred to a more general SP context, but it confirms that social desirability bias may be a problem in f2f surveys. Legget et al. (2003) is a more traditional CV study testing whether the NOAA-panel recommendation to allow f2f 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 payment card 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 a strong 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) compared 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 f2f interviews (as discussed in Section 2.2.1 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 f2f modes. Davis (2004) speculates that this finding is due to time pressure on the telephone (though she also acknowledges that sample composition effects may be important in explaining differences). A recent study by Maguire (2009) also finds higher mean WTP in the telephone mode than the mail and f2f modes, indicating, as the author states, social desirability bias in the telephone mode. Two other studies find mixed results when comparing f2f, phone and self-administered survey modes (Hanley, 1989; Smith, 2006).

In the context of considering choice between self-administered and interview modes, it is also worth pointing out that the SP literature also has documented a range of interviewer effects, in addition to social desirability bias. These are all effects that would be avoided when choosing Internet or mail surveys. Studies investigating interviewer effects in the SP literature include Bateman and Mawby (2004) (impact of interviewer dress appearance), Loureiro and Lotade (2005) (social desirability related to interviewer origin), Mannesto and Loomis (1991) (degree of interviewer experience) and Gong and Aadland (2011) (interviewer gender and race in a telephone survey).

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).[32] 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 these studies), perhaps since such bias has traditionally been related to f2f interviews. Further, many of the studies acknowledge that sample composition effects cannot be distinguished from measurement effects, so they focus more on assessing response rates, data quality in terms of item non-response and other sample biases between modes.[33] They generally have little to say about response quality and satisficing, though two point out that increased time to think in mail surveys (provided such time is actually used) may make mail surveys more suitable for CV questions than telephone surveys (Mannesto and Loomis 1991; Lindberg et al. 1997).[34] This is an interesting area of research also for Internet surveys, e.g. as investigated by Lindhjem and Navrud (2011) for CV and Vista et al. (2009) for CM.

4   DISCUSSION AND IMPLICATIONS FOR SP PRACTICE AND RESEARCH

The use of Internet in Stated Preference (SP) research is increasing rapidly. This review has taken stock of the knowledge of which survey mode effects can be expected, and the experiences to date from studies in the survey methodology and SP literatures that compare Internet with other survey modes. We briefly summarize the main results, before discussing implications for SP practice and research.

---

[32] 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.

[33] A number of studies in the SP literature not reviewed here have investigated in detail such sample effects (especially self-selection and non-response bias) for traditional survey modes: See e.g. Edwards and Anderson (1987) and Whitehead et al. (1993) (telephone and mail), Messonier et al. (2000) (mail and f2f), and Harpman et al.(2004), Mattson and Li (1994), Fredman (1999) and Dalecki et al. (1993) (all mail).

[34] 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.

Net mode effects are the sum of differences in the representation of the target population (coverage, sampling, non-response) and differences in the measurement of the constructs of interest (validity and measurement). Measurement differences between modes have, in the survey methodology literature in particular, been explained by social desirability bias in telephone or face-to-face (f2f) interviews due to the involvement of interviewers and various forms of satisficing (i.e. shortcutting of the response process leading to sub-optimal responses) in self-administered surveys using mail or Internet. Contrary to common belief, and those held by the NOAA panel (Arrow et al. 1993), social desirability bias has in the survey literature typically been found to be higher for telephone than f2f surveys. One reason for this may be the "dampening" effect from the interpersonal trust that can develop in a personal interview setting (as opposed to in a telephone interview). Internet surveys are seen to be similar to mail along this dimension.

Survey research that investigates data quality and degree of satisficing find indications of less satisficing in Internet surveys compared to telephone, similar levels of such effects compared to mail surveys and somewhat higher prevalence compared to f2f surveys. However, research to date is limited and it is not possible to conclude that satisficing generally is a bigger problem or data quality substantially lower for Internet surveys compared to other modes. Hence, results from the general survey literature to date seem to imply that the fear of large measurement effects from Internet surveys may be overblown.

In addition to potential measurement effects, differences between modes may be due to sample composition effects, i.e. *who* responds rather than *how*. Concerns have often been raised about the problems of non-coverage of the general population and high non-response rates for Internet surveys. Although non-coverage is decreasing quickly as Internet penetration goes up, a suitable sampling frame does not exist for Internet surveys. Research

generally shows lower response rates for Internet surveys than for other surveys; and the elderly and people with low education and income are typically underrepresented. Probability and non-probability based Internet panels of willing respondents have been built to deal with both coverage and non-response problems. We found few studies in the survey methodology literature that make stringent comparisons of sample composition effects across different types of Internet panels and other survey modes. There are, however, some indications that probability-based Internet panel surveys are more accurate and reliable, and that post-survey weighting of unrepresentative samples based on demographics may not be a reliable strategy. Even though many samples from non-probability based panels appear representative in terms of socio-economic variables, statistical inference on such samples is unfounded as long as the initial selection is non-random.

The lessons from the survey methodology literature are relevant to SP research since SP surveys share many of the same features common to typical surveys in the other social sciences. In SP research, the primary interest are the questions used to derive welfare measures (WTP or implicit prices), but questions related to use frequencies, attitudes, socio-economic data, and other constructs that can be used to assess validity of the SP data, are also important. 17 Contingent Valuation (CV) and Choice Modelling (CM) studies comparing Internet and other survey modes were identified and reviewed. The goods valued vary widely and cover water and air quality and environment-related health risks, nature protection, landscape amenities, and climate change and energy policies. The studies are generally fairly heterogeneous in design and in the choice of methods and issues compared across modes. Most of the studies utilize Internet panels of some kind. In terms of derived welfare measures, 10 out of 17 pairwise mode comparisons find equality, and 5 find lower values for the Internet survey. Hence, values tend to be conservative compared to other modes. Apart from this, there is no clear evidence that particular modes give higher, similar or lower values; like e.g.

clearly higher values for telephone or f2f modes. However, most of the studies are not able to separate sample composition from measurement effects in a convincing way, and only a couple of studies apply weighting or other post-survey adjustments to try to separate the effects.

The perhaps most well-controlled mode comparison to date for SP, commissioned by U.S. EPA, finds no difference in mean WTP between the mail and Internet surveys, but higher mean WTP for the telephone survey (Grandjean et al. (2009); Taylor et al. (2009)). They attribute this higher WTP to social desirability bias, and speculate that this effect would likely have carried over to a comparison of f2f (which was not included in the study). With the caveat of confounding factors, most studies reviewed find fairly similar degrees of validity and general data quality between Internet and the other survey modes.

What are then the implications of the Internet survey research for SP practice and future research? In terms of research, there are several promising avenues to pursue. First of all, more effort should be put into designing controlled mode comparison experiments, to disentangle both sample composition and measurement effects. There are experiences and methods both from survey methodology and experimental economics that can be fruitfully applied. The challenge here is not just to investigate whether there are statistically significant differences between welfare measures across modes, but to understand how large such differences may be and why they exist. The research to date has only begun to grapple with the first of these challenges. In the practical use of SP results size may matter more than statistical significance. Can welfare measures from Internet surveys and other modes, when generalized over the population, be considered the same (i.e. equivalent[35]) for all practical

---

[35] The analogy of comparing a new, cheaper and more convenient drug with functionally equivalent properties to and old

drug in pharmaceutical research, is quite striking in our case of Internet vs. more expensive, traditional survey modes:

purposes? The answer to this question will of course depend on the level of difference deemed acceptable for passing convergent validity considerations for survey modes.

To better understand why mode effects exist, much can be learnt from psychology and existing survey mode research. However, economics can also further develop its own contributions. 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". This is a constrained optimization problem economists are used to analyze. Hence, as more and more complex and sophisticated Internet surveys are developed and fielded, research has not really taken on the challenge put forward by Kerry Smith of developing a theoretical framework to understand survey behavior in SP research. The studies we have reviewed mostly grapple with empirical effects, and have little in the way of theoretical or conceptual basis for developing consistent hypotheses that can be stringently tested. However, there are areas of SP research that may be fruitfully expanded to more explicitly address measurement effects. For example, the research on measuring and adjusting for stated respondent uncertainty using different response formats (see e.g. Olsen 2011, Shaikh et al. 2007) seems to be related to

---

"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). Equivalency tests have also been used to test the validity of benefit transfer of SP survey results (Kristofferson and Navrud 2005)

satisficing[36]: Higher respondent uncertainty could indicate a higher degree of satisficing behavior (though some of the uncertainty may not be reduced through greater respondent effort). However, higher stated certainty may not (always) mean higher quality responses (e.g. respondents could express high certainty about a socially desirable response or a response following little effort). To better understand the underlying mechanisms of relevance to respondent behavior, the sources of the uncertainty would need to be better understood. Further work to understand respondent behavior and mode effects needs to be interdisciplinary and Jason Shogren's general warning to experimental economists is pertinent here: "economists venturing into this cognitive minefield alone will end up fifty years behind the psychologist's times" (Shogren, 2005).

Internet surveys open up a methodologically interesting, cheap and convenient medium for further experimental research to understand people's preferences and reduce behavioral anomalies. Visual stimuli, rather than words and numerical information, are increasingly seen as the most effective way to promote comprehension (Bateman et al. 2009). In this regard Internet surveys provide almost limitless opportunities, as broadband and fiber cable capacities quickly increase. For example, in much of the Western world the transformation of energy systems from fossil fuels to renewable energy will give rise to large landscape changes that could be much better visualized and conveyed using video[37] or virtual reality in an Internet application. Other promising avenues of research for Internet surveys include more stringent measurement of the impact of time use, comprehension and complexity and satisficing effects and investigation of impacts of survey incentives (that many Internet panels use) for participation and response. It would also be fruitful to investigate whether respondents generally find (innovative) SP Internet surveys more interesting and enjoyable

---

[36] We thank an anonymous reviewer for this point.

[37] See Navrud (1997) for an early and rare application of video in f2f CV studies.

than the more standard marketing surveys they typically face, and if so, whether they spend more time and effort potentially giving lower satisficing effects in SP surveys.[38] Recently in Internet survey research, more interactive and personalized surveys that can enhance learning effects have been developed. There is considerable scope also in SP surveys to utilize such techniques to encourage higher response, learning through the survey and better response quality. However, trends in Internet surveys towards more human-like computer surveys (using voice, visuals, video etc.) may also introduce social desirability bias or generally less honest responding also for Internet surveys (Couper 2005). These are all worthy topics of further investigation.

However, while such experimentation will go on and will no-doubt be very useful to SP research, environmental economists will also have to deal with the more mundane challenges of using Internet surveys to derive reliable welfare measures for a general population. This means going beyond split-sample experimentation on non-probability based samples. In this respect, our review has also shown that there is currently a knowledge gap in SP practice and research that needs to be filled. Most papers do not distinguish between panels that recruit randomly or through convenience.[39] It is important to understand that differences in composition and practices of individual panels can influence survey results (Baker et al 2010). It seems in many cases that too much of the data collection process is handed to Internet survey companies that often operate in a "black box", in terms of recruitment strategies, quota sampling to resemble representativity etc. This means that the real (not just final stage) response rates and the error structure of the welfare estimates are often unknown. It also means that the significance of non-response biases is hard to judge. Further research on the

---

[38] We thank Peter Boxall for pointing this out.

[39] This is of course also important if an experiment is meant to compare two survey modes only, rather than for deriving welfare measures and generalizing to a population.

extent of non-response bias in SP Internet surveys is clearly needed. More stringency and transparency in the treatment of sampling and generalization, including understanding the impacts of post-survey weighting strategies, would also be required to give credibility to derived welfare measures. Finally, despite some of these current weaknesses of Internet panel research, the panels also contain useful background information of survey respondents that may give us a way to better understand differences between respondents and non-respondents and potential non-response biases.

In terms of general guidance for current SP practitioners regarding the use of Internet surveys and panels, there are no hard and fast rules that are likely to apply to all types of Internet surveys. This is of course partly because different surveys have different purposes, and partly because it is still early days for Internet survey research. However, for SP researchers who contemplate the use of an Internet panel, Baker et al. (2010) in their Appendix B provides a useful list of questions researchers should ask of potential such panels, and a description of why the questions are important. As the research and practices using Internet surveys develop in the SP literature, it should perhaps be an ambition to provide specific guidance and best practice suggestions that could apply to different types of SP surveys, especially as these are increasingly used to inform cost-benefit analyses and policy decisions.

5   REFERENCES

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

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

Atkeson, R. L., Adams, A. L., Bryant, L. A., Zilberman, L. and Saunders, K. L., 2011. Considering Mixed Mode Surveys for Questions in Political Behavior: Using the Internet and Mail to Get Quality Data at Reasonable Costs. Political Behavior 33(1); 161-178.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K. and Zahs, D., 2010. Research Synthesis: AAPOR report on online panels. Public Opinion Quarterly, 74(4); 711-781.

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

Barrio, M. and Loureiro, M., 2010. A meta-analysis of contingent valuation forest studies. Ecological Economics, 69(5); 1023-1030.

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

Bateman, I. J., Day, B. H., Jones, A. P. and Jude, S., 2009. Reducing gain–loss asymmetry:Avirtual reality choice experiment valuing land use change. Journal of Environmental Economics and Management, 58; 106-118.

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

Bell, J., Huber, J. and Viscusi, W. K., 2011. Survey Mode Effects on Valuation of Environmental Goods. International Journal of Environmental Research and Public Health, 8(4); 1222-1243.

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

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

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

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

Borkan, B., 2010. The Mode Effect in Mixed-Mode Surveys Mail and Web Surveys. Social Science Computer Review, 28(3); 371-380.

Boxall, P., Adamowicz, W. L. and Moon, A., 2009. Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement. Australian Journal of Agricultural and Resource Economics, 53(4); 503-519.

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

Brown, G. and Hagen, D. A., 2010. Special Issue: Behavioral Economics and the Environment. Environmental and Resource Economics, 46(2); 135-247.

Cai, B. L., Cameron, T. A. and Gerdes, G. R., 2010. Distributional Preferences and the Incidence of Costs and Benefits in Climate Change Policy. Environmental & Resource Economics, 46(4); 429-458.

Cameron, T. A., Poe, G. L., Ethier, R. G. and Schulze, W. D., 2002. Alternative non-market value-elicitation methods: Are the underlying preferences the same? Journal of Environmental Economics and Management, 44(3); 391-425.

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

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

Carlsson, F., Kataria, M. and Lampi, E., 2010. Dealing with Ignored Attributes in Choice Experiments on Valuation of Sweden's Environmental Quality Objectives. Environmental and Resource Economics, 47; 65-89.

Chang, L. and Krosnick, J. A., 2009. National Surveys Via Rdd Telephone Interviewing Versus the Internet. Public Opinion Quarterly, 73(4); 641-678.

Chang, L. and Krosnick, J. A., 2010. Comparing Oral Interviewing with Self-Administered Computerized QuestionnairesAn Experiment. Public Opinion Quarterly, 74(1); 154-167.

Cooper, M., 2008. Designing Effective Web Surveys Cambridge University Press, pp.

Couper, M. and Miller, P. V., 2008. Special Issue: Web Survey Methods. Public Opinion Quarterly, 72(5); 831-1032.

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

Couper, M. P., Kapteyn, A., Schonlau, M. and Winter, J., 2007. Noncoverage and nonresponse in an Internet survey Social Science Research, 36(1); 131-148.

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

Dalecki, M. G., Whitehead, J. C. and Blomquist, G. C., 1993. Sample Nonresponse Bias And Aggregate Benefits In Contingent Valuation - An Examination Of Early, Late And Non-Respondents. Journal of Environmental Management, 38(2); 133-143.

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

de Leeuw, E. D., 2005. To Mix or Not to Mix Data Collection Modes in Surveys. Journal of Official Statistics, 21(2); 233-255.

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

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.

Denscombe, M., 2009. Item non-response rates: a comparison of online and paper questionnaires. International Journal of Social Research Methodology, 12(4); 281-291.

DeShazo, J. R. and Fermo, G., 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., Gerking, S. and Goffe, W. L., 2007. Valuation of Non-Market Goods Using Computer-Assisted Surveys: A Comparison of Data Quality from Internet and RDD Samples

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

Dillman, D. and Bowker, J. M., 2001. The WEB questionnaire challenge to survey methodologists, in U.-D. Reips and M. Bosnjak, Eds, Dimensions of Internet Science. Pabst Science Publishers,Lengerich, Germany, 159-178.

Dillman, D., Smyth, J. D. and Christian, L. M., 2009. Internet, Mail and Mixed-Mode Surveys. Wiley, pp.

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

Fleming, C. M. and Bowden, M., 2009. Web-based surveys as an alternative to traditional mail methods. Journal of Environmental Management, 90; 284-292.

Fredman, P., 1999. A test of nonresponse bias in a mail contingent valuation survey, in M. Boman, et al., Eds, Topics in environmental economics. Kluwer Academic Publishers, 175-186.

Fricker, S., Galesic, M., Tourangeau, R. and Yan, T., 2005. An experimental comparison of web and telephone surveys. Public Opinion Quarterly, 69(3); 370-392.

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

Galesic, M., Tourangeau, R., Couper, M. P. and Conrad, F. G., 2008. Eye-Tracking Data New Insights On Response Order Effects And Other Cognitive Shortcuts In Survey Responding. Public Opinion Quarterly, 72(5); 892-913.

Gong, M. and Aadland, D., 2011. Interview Effects in an Environmental Phone Survey. Environmental and Resource Economics, 49(1); 47-64.

Grandjean, B. D., Nelson, N. M. and Taylor, P. A., 2009. Comparing an Internet Panel Survey to Mail and Phone Surveys on Willingness to Pay for Environmental Quality: A National Mode Test. Paper presented at the 64th Annual conference of the American Association for Public Opinion Research, 14-17 May 2009.

Green, C. and Tunstall, S., 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., 2006. Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly, 70(5); 646-675.

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

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

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

Heerwegh, D., 2009. Mode Differences Between Face-To-Face And Web Surveys: An Experimental Investigation Of Data Quality And Social Desirability Effects. International Journal of Public Opinion Research, 21(1); 111-121.

Heerwegh, D. and Loosveldt, G., 2008. Face-To-Face Versus Web Surveying In A High-Internet-Coverage Population Differences In Response Quality. Public Opinion Quarterly, 72(5); 836-846.

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

Holbrook, A. L. and Krosnick, J. A., 2010. Social desirability bias in voter turnout reports. Public Opinion Quarterly, 74(1); 37-67.

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

Hynes, S. and Hanley, N., 2006. Preservation versus development on Irish rivers: whitewater kayaking and hydro-power in Ireland. Land Use Policy, 23; 170-180.

Internet World Stats, 2010. Internet Usage in European Union: http://www.internetworldstats.com/stats9.htm

Jäckle, A., Roberts, C. and Lynn, P., 2010. Assessing the Effect of Data Collection Mode on Measurement. International Statistical Review, 78(1); 3-20.

Kramer, R. A. and Eisen-Hecht, J. I., 2002. Estimating the economic value of water quality protection in the Catawba River basin. Water Resources Research, 38(9); 1-10.

Kreuter, F., Presser, S. and Tourangeau, R., 2008. Social Desirability Bias In Cati, Ivr, And Web Surveys The Effects Of Mode And Question Sensitivity. Public Opinion Quarterly, 72(5); 847-865.

Kristofersson, D. and S. Navrud 2005: Validity Tests of Benefit Transfer – Are We Performing the Wrong Test? Environmental and Resource Economics 30 (3); 279-286.

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

Ladenburg, J. and Olsen, S. B., 2008. Gender-specific starting point bias in choice experiments: Evidence from an empirical study. Journal of Environmental Economics and Management, 56(3); 275-285.

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

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

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

Lienhoop, N. and Fischer, A., 2009. Can you be bothered? The role of participant motivation in the valuation of species conservation measures. Journal of Environmental Planning and Management, 52(4); 519-534.

Lindberg, K., Johnson, R. L. and Berrens, R. P., 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 Navrud, S., 2009. Asking for Individual or Household Willingness to Pay for Environmental Goods: Implication for aggregate welfare measures Environmental and Resource Economics, 43(1); 11-29.

Lindhjem, H. and Navrud, S. (2011). Can panel-based Internet surveys supplement costly in-person interviews in contingent valuation of environmental goods? Ecological Economics, 70(9); 1628-1637..

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

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

Loomis, J., Miller, J., Gonzales-Caban, A. and Champ, P. A., 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 Lotade, J., 2005. Interviewer Effects on the Valuation of Goods with Ethical and Environmental Attributes. Environmental & Resource economics, 30; 49-72.

MacDonald, D. H., Morrison, M., Rose, J. and Boyle, K., 2010. Untangling Differences in Values from Internet and Mail Stated Preference Studies. Paper presented at the World Congress of Environmental and Resource Economists, Montreal, Canada, June 28 - July 2, 2010.

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

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

Malhotra, N., 2009. Order Effects in Complex and Simple Tasks. Public Opinion Quarterly, 73(1); 180-198.

Malhotra, N. and Krosnick, J., 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.

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

Mannesto, G. and Loomis, J. B., 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., Freitas, H. and Domingos, T., 2007. Testing for the survey mode effect on contingent valuation data quality: a case study of web based versus in-person interviews. Ecological Economics, 62; 388-398.

Mathews, K. E., Freeman, M. L. and Desvouges, W. H., 2006. How and How Much? The Role of Information in Stated Choice Questionnaires, in B. Kanninen, Eds, Valuing Environmental Amenities Using Stated Choice Studies. Springer.

Mattsson, L. and Li, C. Z., 1994. Sample Nonresponse In A Mail Contingent Valuation Survey - An Empirical-Test Of The Effect On Value Inference. Journal of Leisure Research, 26(2); 182-188.

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

Meyerhoff, J. and Liebe, U., 2009. Status Quo Effect in Choice Experiments: Empirical Evidence on Attitudes and Choice Task Complexity. Land Economics, 85(3); 515-528.

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

Navrud, S., 1997. Communciation Devices in Contingent Valuation Surveys - Experiments with video, in Kopp. R.J., W.-W. Pommerehne and N. Schwartz, Eds, Determining the Value

of Non-marketed Goods: Economics, Psychological and Policy Relevant Aspects of Contingent Valuation. Kluwer Academic Publishers, Holland.

Navrud, S. and Pruckner, G., 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.

Nielsen, J. S., 2011. Use of the Internet for willingness-to-pay surveys. A comparison of face-to-face and web-based interviews. Resource and Energy Economics, 33(1); 119-129.

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

Olsen, S. B., Lundhede, T. H., Jacobsen, J. B. and Thorsen, B. J. 2011. Tough and Easy Choices: Testing the Influence of Utility Difference on Stated Certainty-in-Choice in Choice Experiments. Environmental and Resource Economics, 49(4); 491-510.

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

Rookey, B. D., Hanway, S. and Dillman, D. A., 2008. Does A Probability-Based Household Panel Benefit From Assignment To Postal Response As An Alternative To Internet-Only? Public Opinion Quarterly, 72(5); 962-984.

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

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

Schulze, W., McClelland, G., Waldman, D. and Lazo, J. H., 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 Strasman, T. J., 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.

Shaikh, S. L., Sun, L. and van. Kooten, G. C. (2007). Treating respondent uncertainty in contingent valuation: A comparison of empirical treatments. Ecological Economics, 62(1); 115-125.

Shih, T. H. and Fan, X. T., 2008. Comparing response rates from Web and mail surveys: A meta-analysis. Field Methods, 20(3); 249-271.

Shogren, J. F., 2005. Experimental methods and valuation, in K.-G. Maler and J. R. Vincent, Eds, Handbook of Environmental Economics. North Holland,Amsterdam.

Simon, H. A., 1956. Rational choice and the structure of the environment. Psychological Review, 63(2); 129-138.

Smith, R. D., 2006. It's not just *what* you do, it's the *way* you do it: the effect of different payment card formats and survey administration on willingness to pay for health gains. Health Economics, 15; 281-293.

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

Stocke, V. and Hunkler, C., 2007. Measures of desirability beliefs and their validity as indicators for socially desirable responding Field Methods 19(3); 313-336.

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

http://yosemite1.epa.gov/ee/epa/eerm.nsf/Author/A62D95F235503D03852575A800674D75, USEPA,

Thurston, H. W., 2006. Non-market valuation on the internet, in A. Alberini and J. Kahn, Eds, Handbook on contingent valuation. Edward Elgar.

Toepoel, V., Vis, C., Das, M. and van Soest, A., 2009. Design of Web Questionnaires An Information-Processing Perspective for the Effect of Response Categories. Sociological Methods & Research, 37(3); 371-392.

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

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

Tourangeau, R., Groves, R. M., Kennedy, C. and Yan, T., 2009. The Presentation of a Web Survey, Nonresponse and Measurement Error among Members of Web Panel. Journal of Official Statistics, 25(3); 299-321.

Tourangeau, R., Groves, R. M. and Redline, C. D., 2010. Sensitive Topics And Reluctant Respondents Demonstrating A Link Between Nonresponse Bias And Measurement Error. Public Opinion Quarterly, 74(3); 413-432.

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

Tsuge, T. and Washida, T., 2003. Economic valuation of the Seto Inland Sea by using an Internet CV survey. Marine Pollution Bulletin, 47; 230-236.

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

Vista, A. B., Rosenberger, R. S. and Collins, A. R., 2009. If you provide it, will they read it? Response time effects in a choice experiment. Canadian Journal of Agricultural Economics, 57(3); 365-377.

Whitehead, J. C., Groothuis, P. A. and Blomquist, G. C., 1993. Testing For Nonresponse And Sample Selection Bias In Contingent Valuation - Analysis Of A Combination Phone Mail Survey. Economics Letters, 41(2); 215-220.

Whittaker, D., Vaske, J. J., Donnelly, M. P. and DeRuiter, D. S., 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., Smith, V. K., Okorafor, A., Okore, A., Liu, J. L. and McPhail, A., 1992. Giving Respondents Time To Think In Contingent Valuation Studies - A Developing-Country Application. J.Environ.Econ.Manage., 22(3); 205-225.

Windle, J. and Rolfe, J., 2011. Comparing Responses from Internet and Paper-Based Collection Methods in more Complex Stated Preference Environmental Surveys. Economic Analysis & Policy, 41(1); 83-97.

Yeager, D. S., Krosnick, J., Chang, L., Javitz, H. S., Levindusky, M. S., Simpser, A. and Wang, R., 2009. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples, Stanford University