Roberts, Caroline; Jäckle, Annette

Working Paper

Causes of mode effects: Separating out interviewer and stimulus effects in comparisons of face-to-face and telephone surveys

ISER Working Paper Series, No. 2012-27

Provided in Cooperation with:
Institute for Social and Economic Research (ISER), University of Essex

Suggested Citation: Roberts, Caroline; Jäckle, Annette (2012) : Causes of mode effects: Separating out interviewer and stimulus effects in comparisons of face-to-face and telephone surveys, ISER Working Paper Series, No. 2012-27, University of Essex, Institute for Social and Economic Research (ISER), Colchester

This Version is available at:
https://hdl.handle.net/10419/91682

ISER Working Paper Series

Causes of Mode Effects: Separating Out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys

Caroline Roberts
Institute of Social Sciences
University of Lausanne

Annette Jäckle
Institute for Social and Economic Research
University of Essex

No. 2012-27
November 2012

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-technical summary

Many long-established surveys that have so far relied on face-to-face interviewing are facing increasing pressures to reduce data collection costs. As a result, there is rising interest in the use of cheaper data collection methods, such as telephone interviewing, alongside or instead of face-to-face interviews. For established surveys that produce time series or longitudinal data, there is concern that changing the mode of data collection may pose a threat to the continuity of the data produced. In this paper we focus on reasons why respondents may answer a given survey question differently depending on whether they are interviewed face-to-face or over the telephone. The main differences between telephone and face-to-face interviewing are in the physical presence of the interviewer, and the aural versus visual presentation of questions. Previous studies that have examined whether telephone interviewing leads respondent to give different answers than face-to-face interviewing have tended to confound these two differences between the modes, making it difficult for survey designers to identify the causes of differences and how the survey design might be adapted to reduce risks to data comparability.

In this paper we use data from an experiment conducted in the context of the European Social Survey to test for mechanisms that may lead respondents to give different answers in telephone and face-to-face interviewing. The experiment included three randomly allocated treatment groups: (1) face-to-face interviewing using showcards to present response categories visually, (2) telephone interviewing, and (3) face-to-face interviewing using the telephone questionnaire, that is without showcards.

The results suggest that measurement differences between telephone and face-to-face interviewing with showcards were mainly driven by the physical presence of the interviewer: telephone respondents were more likely to give socially desirable responses than face-to-face interviewers. The question stimulus, that is, whether questions were administered aurally or with the help of showcards, did not affect responses.

Causes of Mode Effects: Separating Out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys

Caroline Roberts (University of Lausanne)

Annette Jäckle (University of Essex)

Abstract: We attempt to isolate the causes of mode effects on measurement in a comparison of face-to-face and telephone interviewing, distinguishing between effects caused by differences in the type of question stimulus used in each mode (audio vs. visual) and effects caused by other differences between the modes, notably, the presence or absence of the interviewer. We use data from an experiment conducted in the context of the European Social Survey. Differences in the stimulus did not lead to differential measurement error, but the presence or absence of the interviewer did. Telephone respondents were far more likely to give socially desirable responses than face-to-face respondents.

Keywords: showcards, measurement, satisficing, social desirability bias

JEL classification: C83

Contact: Caroline Roberts, Institute of Social Sciences, University of Lausanne, Room 5548, Bâtiment Géopolis, 1015 Lausanne, Switzerland, Tel: +41 (0)21 692 38 39, email: caroline.roberts@unil.ch

Acknowledgements: The data used in this study were collected as part of the European Social Survey's programme of methodological research on mixed mode data collection, co-funded by the European Commission and Gallup Europe. The authors wish to thank Peter Lynn for his substantial contribution to the design of the experiment, and Robert Manchin, Agnes Illyes and Gergely Hideg, who directed the data collection and data management at Gallup Europe.

1 Introduction

Many long-established, high quality surveys that have traditionally relied on face-to-face interviewing are facing pressure to reduce the cost of their fieldwork. This has led to a growing interest in cheaper data collection alternatives, which increasingly imply the use of a mix of survey modes to ensure adequate population coverage. For example, telephone interviewing is one of the most viable alternatives to face-to-face interviewing for large-scale surveys of the general population, offering fast, convenient and low-cost data collection. The growing problem of under-coverage due to the rapid and widespread take-up of mobile phones (e.g. Peytchev, Carley-Baxter and Black, 2011), however, increasingly limits its use as a stand-alone mode. While mixed mode surveys are quickly becoming common practice in the commercial world, for government or academically-led surveys that produce time series or longitudinal data the prospect of switching or mixing modes poses a threat to the continuity of the estimates they produce. In this study we examine whether and why responses may differ when telephone interviewing is used alongside face-to-face interviewing.

The mode choice has important implications for the quality of the data produced by a survey, influencing the size and nature of non-sampling errors, and how they inter-relate. Notably, mode can influence who is able to participate in a survey (coverage), who decides to participate (non-response) and how the questions are answered (the quality of measurement). So, if a survey administers the same questions to different respondents in different modes (de Leeuw 2005), there is a risk that each mode's unique composition of errors will render the data incomparable. By confounding different sources of error, mixing modes of data collection can seriously impact the usability of the data, so the need to be able to measure and control for different types of effect separately has become paramount in complex multimode survey designs (Vannieuwenhuyze, Loosveldt, and Molenberghs 2010; Buelens et al., 2012).

So that appropriate measures can be taken either to *prevent* mode effects from limiting the equivalence of the data (e.g. such as in Dillman's (2007) 'unimode' questionnaire construction approach) or to *correct* for problems at the analysis stage (Couper 2011; p. 899), identifying and understanding the causes of differences observed in mixed mode data has become a priority for practitioners wishing to combine modes. From a Total Survey Error perspective (Biemer 2010; Groves and Lyberg 2010) it is particularly important to assess the inter-play between the major forms of mode effect in mixed mode surveys (the combined effect of selection and measurement errors being perhaps the most important), so that informed conclusions can be drawn about the relative advantages and disadvantages of mixed

mode designs for survey errors and costs (Groves 1989).

In the present study, we focus uniquely on the factors contributing to differential measurement errors between modes, while attempting to control for selection effects through the sample design and multivariate analytic techniques. Differential measurement error remains a unique challenge on its own, because it is not always possible to predict when differences will occur, whether differences will matter in terms of their likely effect on the conclusions made by analysts, and if so, how they might be mitigated (Jäckle, Roberts and Lynn 2010). Although previous studies have tested for differences in measurement between telephone and face-to-face interviewing, their ability to infer the likely causes of mode effects was often limited (see Holbrook, Green and Krosnick, 2003). As a result previous studies provide only limited information about how best to enhance data comparability when combining telephone and face-to-face interviewing.

We examine the causes of measurement differences between telephone and face-to-face interviewing, using experimental data related to the European Social Survey (ESS). We examine the roles that the interviewer and the nature of the question stimulus play in causing mode effects. In Section 2 we review the literature on why respondents may answer a given question differently if interviewed by telephone or face-to-face. Section 3 sets out the hypotheses we test; Section 4 describes the data and analysis methods; Section 5 discusses the results and Section 6 provides an overall discussion and conclusion.

2 Causes of measurement differences in telephone and face-to-face interviewing

The ways in which telephone and face-to-face interviewing can influence data quality have already received considerable attention in the literature on mode effects, during an earlier wave of interest when telephone interviewing first gained popularity in survey research during the 1970s. At this time, Groves (1979) made the distinction between measurement error arising from the behavior of the different 'actors' involved in the data collection process (notably, the interviewer and the respondent), and error arising from the 'questions' used to collect the data (the design of the questionnaire and the way in which it is administered), arguing the need to focus on any changes to either required by a switch in interviewing mode in order to preempt measurement differences (ibid, p. 190). In the following, we discuss these two sources of influence on data quality.

Mode effects arising from interviewer and respondent interaction

Telephone and face-to-face interaction differ in terms of the *sensory channel* available for the

transmission of information and the *physical presence* of the interviewer (de Leeuw 2005). Telephone interviews are restricted in the types of 'messages' that can be transmitted between participants (Groves 1990). By contrast, face-to-face interviews can make use of a wide range of visual cues (such as facial gestures and body language) to communicate a variety of information that cannot be conveyed verbally. For example, interviewers are able to use nonverbal cues to detect whether respondents are having difficulty understanding certain questions or losing interest in the interview (ibid, p. 83), and non-verbal cues can help to fill gaps in conversation – such as when respondents are thinking about their answers. Audio-only communication is more restricted. As a result, telephone interviewers may be less able to influence the motivation and command the attention of respondents, meaning respondents may be more likely to be engaged in other activities while answering survey questions. They may also be inclined to speed up the pace of the interview, in order to minimize awkward silences (Groves 1989), with the result that telephone interviews tend to be shorter than face-to-face interviews.

These fundamental characteristics of telephone and face-to-face interviewing have been argued to influence *respondent motivation,* as well as the *cognitive burden* of the survey task (Holbrook, Green and Krosnick 2003). Both motivation and task difficulty are important factors influencing the quality of respondents' answers (Krosnick 1991), because they, respectively, affect the amount of effort respondents make and the amount of effort required to answer a survey question. Cognitive models of the survey response process (e.g. Cannell, Miller, and Oksenberg 1981; Tourangeau, Rips, and Rasinski 2000) rest on the assumption that under optimal conditions respondents make sufficient effort to systematically execute the necessary components of processing involved in answering questions (comprehension of the question, retrieval of relevant information from memory, use of that information to make required judgments, and selection and reporting of an answer, ibid; p.7). According to the theory of satisficing (Krosnick 1991), respondents who are not sufficiently motivated or unable to make sufficient effort, will either complete each component of processing less thoroughly (in the case of so-called 'weak satisficing') or skip one or more of the stages altogether (in the case of 'strong satisficing'), to provide merely satisfactory, rather than optimal responses (p.215). Consistent with the theory, Holbrook, Green and Krosnick (2003) show that respondents in telephone interviews are more likely to give acquiescent answers (to agree to questions irrespective of their content), to give 'no opinion' answers, and to give identical answers to items rated on the same scale than respondents in face-to-face interviews.

Furthermore, respondents with lower cognitive ability interviewed by telephone are particularly likely to reveal satisficing errors in their answers (ibid.).

The physical presence of the interviewer and the availability of nonverbal cues also play a role in the build-up of *rapport* between interviewers and respondents in face-to-face interviews, making in-person interaction more intimate than that conducted by telephone (Groves and Kahn 1979). Other factors are likely to be important too, such as the social norms surrounding interactions with strangers in person compared to on the telephone, which likely influence how the relationship between interviewer and respondent is defined at the outset (Groves 1990). The enhanced intimacy of interactions in face-to-face mode can have consequences for the honesty with which respondents are willing to report their behaviours and attitudes, especially those of a sensitive nature. Indeed, there is mounting evidence that the better rapport established in face-to-face interviewing and the greater opportunities it provides for reassuring respondents of the credibility of the survey and confidentiality of the data, makes it a more effective method than telephone interviewing for obtaining potentially sensitive information (e.g. Smith 1984; de Leeuw and van der Zouwen 1988; Holbrook, Green and Krosnick 2003).

While response errors associated with satisficing and social desirability bias are not the only forms of measurement error likely to differentially affect face-to-face and telephone interviews, they represent two important sources of error resulting from the way in which mode characteristics influence the behaviour of interviewers and respondents (Holbrook, Green and Krosnick 2003).

Mode effects arising from the design of questionnaires

A further advantage of being able to make use of visual channels of communication is the possibility of using showcards to display the available response options to the respondent while the interviewer reads out the questions. Showcards have long been used in face-to-face interviewing with the aim of reducing the cognitive burden on respondents. For example, they are used as prompts to remind respondents of the available response alternatives, particularly where the list of alternatives is long or where a battery of items relies on the same response scale (Miller 1984; Sykes and Collins 1988). They can also be used to assist respondent recall of historical events, and as a method of enhancing the privacy of the response process (Duffy 2003). However, although some interviewers may find showcards useful (Rogers 1976; Jordan, Marcus, and Reeder 1980), the evidence that they facilitate the response process for the respondent is, in fact, rather limited (Miller 1984). Some studies

suggest that showcards may actually increase the burden on respondents, who may be expected to read and absorb the information presented on the card in a relatively short space of time (Sykes and Collins 1988; Duffy 2003), and may find it distracting to listen to the interviewer at the same time as reading the card (Dijkstra and Ongena 2002). Furthermore, the visual layout of information on the showcard may bias response selection, producing response order effects (Krosnick and Alwin 1987; Duffy 2003), or over-reporting of labeled categories (Groves 1990). If the response alternatives are presented visually, then according to Krosnick (1991), the respondent may abandon processing near the start of the list as soon as a satisfactory answer can be selected, resulting in primacy effects (p. 216). If the interviewer reads out the response alternatives, then the effect of response order is less easy to predict. On the one hand, less time is available for processing the earlier options, so the alternatives near the end of a list may receive most attention, resulting in recency effects. On the other hand, weak satisficers may start processing the earliest items and give up because they are unable to retain all the options in memory as the list of alternatives grows, resulting in primacy effects.

If visual aids in survey interviews influence the amount of effort needed to answer questions, or encourage respondents to favour particular response alternatives over others, then questions with showcards are likely to be susceptible to errors that other questions, administered solely visually or orally may not be. This seems even more plausible given that showcard questions are often longer and more complex than questions that can be asked without them (Smith 1987), and so already place considerable cognitive demands on the respondent (Wikman and Waerneryd 1990). The unique mix of measurement errors on these types of questions in face-to-face surveys poses a challenge to researchers seeking to collect the same data in another mode. Indeed, in comparisons between face-to-face and telephone interviews, questions with showcards in face-to-face interviews have been shown to elicit different answers from respondents when administered by telephone (Groves 1990; Groves and Kahn 1979). However, it is not always clear whether this is due to the unique measurement properties of the mode, or the fact that showcard questions simply cannot be asked in an identical way when a survey is conducted in other modes, meaning that respondents are exposed to different question stimuli. Correspondingly, Dillman (2000) has argued that, "the most basic cause of mode differences is the tendency for people to construct questions differently for two types of questionnaire" (p.224). Confounds of this type and others are not uncommon in the literature on mode effects, in studies seeking to identify the

effect of mode on the comparability of measurement. Holbrook, Green and Krosnick (2003), for example, reviewed studies involving face-to-face and telephone comparisons to assess the available evidence relating to the prevalence of satisficing and social desirability bias in each mode. More than half of the studies they reviewed had confounded mode comparisons; the remainder were limited either due to differences in the questionnaires, or the fact that respondents were not randomly assigned to mode, thereby confounding differential measurement errors across modes with possible selection biases. The authors' own experiments were limited in that they excluded from their analyses all items for which a showcard was used in the face-to-face data collection, therefore limiting their analyses to items that might be easier to adapt for telephone interviewing. Mode effects can result from a variety of influences on measurement quality, and identifying the specific causes of a given effect may provide insight into how mode differences might be mitigated.

3 Hypotheses

The preceding review highlighted different ways in which respondents' answers may be influenced in face-to-face and telephone interviews, which may result in differences in measurement when the two modes are combined. In short, modes are likely to lead to differences in measurement if they have differential effects on the ways in which respondents' execute the cognitive processes involved in answering questions. By influencing the amount of effort respondents must expend to answer questions, their motivation to do so, as well as perceptions about the degree of anonymity with which they can report their answers, the characteristics of each mode affect the prevalence of errors arising from shortcutting the cognitive processes in survey response (satisficing) and social desirability bias (Holbrook, Green and Krosnick, 2003). This gives rise to the following specific hypotheses regarding the causes of mode effects in face-to-face and telephone interviews. ***H1*** and ***H2*** relate to differences between the modes due to differences in the proximity of the interviewer; ***H3*** and ***H4*** relate to differences arising from the nature of the question stimulus:

H1: We expect to see less shortcutting with face-to-face interviewing (without showcards) than with telephone interviewing. This difference between the two modes may occur through two different mechanisms. Firstly the interviewer's presence in face-to-face interviews is expected to reduce the cognitive burden on respondents, by making use of a range of communication channels to facilitate the respondents' comprehension of the survey task, and moderating the pace of the survey interview. Second, the interviewer's presence is

expected to increase respondent motivation, because the respondent can observe nonverbal cues of the interviewer's commitment and enthusiasm, while the interviewer can detect nonverbal cues of declining motivation and react to these. Both mechanisms would produce less shortcutting in face-to-face than telephone interviewing.

Following satisficing theory, we expect the impact of mode to be largest among respondents with low cognitive ability, because increased task difficulty and decreased motivation are likely to be more detrimental for these respondents. If this is true, an extension to ***H1*** is that we should see larger differences in the extent of shortcutting across modes for low ability respondents than for high ability respondents.

H2: We expect the extent of social desirability bias to differ between face-to-face and telephone interviewing. The direction of this difference is not a priori clear. On the one hand, the interviewer's presence reduces anonymity and 'social distance' and may make the reporting of sensitive information more threatening for the respondent. Fear of sanctions (such as nonverbal signs of approval or disapproval from the interviewer) is likely to reduce respondents' willingness to disclose sensitive or potentially embarrassing information. If this is true then we should see more socially desirable answers with face-to-face interviewing. On the other hand, the interviewer's presence improves the rapport with the respondent, because nonverbal communication aids the development of interpersonal trust. In addition, interviewers may be better able to convey the legitimacy of the survey in person than over the telephone. In comparison, telephone respondents may feel less confident that the interviewer will protect the confidentiality of their responses and so be less willing to disclose personal information. If this is true then we should see more socially desirable answers with telephone interviewing.

H3: We expect the extent of shortcutting to be different with showcards than without. The direction of this difference is however not clear a priori. On the one hand, showcards are used with the aim of simplifying the response task, because the visual presentation reduces the burden on the respondent to remember response categories and may make it easier to understand the question. If this is true, then we should see less shortcutting with showcards than without. On the other hand, showcards may unintentionally increase the cognitive burden on respondents, who have to read information in addition to listening to the interviewer. If this is true, then we should expect more shortcutting with showcards than without.

H4: The responses produced by shortcutting are likely to be different with aural and

visual presentations. With showcards respondents are likely to read down the list until they find a plausible answer. With aural presentation respondents are more likely to remember the last response categories. If there is shortcutting, we should therefore expect to see more responses from earlier categories ('primacy effect') with showcards and more responses from later categories ('recency effect') without showcards.

4 Research design

The present study is based on data from a survey experiment conducted in Hungary in July 2005. The experiment formed part of an on-going methodological programme of research designed to inform decisions by the co-ordinating team of the European Social Survey (ESS) on whether to change its policy of single-mode data collection using face-to-face interviews to a mixed mode data collection strategy, and if so, which modes to mix and how (for a discussion of the issues faced by the ESS, see Jäckle, Roberts, and Lynn 2006 and Martin 2011). The objective was to compare data from face-to-face survey interviews with data from telephone interviews, controlling for how questions are asked in each mode. In particular, we wished to distinguish between what we refer to as 1) 'stimulus effects', resulting from differences in the question form or medium in which the response categories are communicated (e.g. whether or not showcards are used); and 2) mode effects per se, resulting from other features of the mode – notably, the presence or absence of the interviewer, and resulting aspects such as the pace with which the interview is conducted and the impersonality, legitimacy and cognitive burden imposed by each mode (Tourangeau, Rips and Rasinski 2000; pp.290-293). The experiment therefore included three treatment groups:

(1) Face-to-face interview with showcards.

(2) Telephone interview with the Group 1 questionnaire adapted for oral administration.

(3) Face-to-face interview without showcards using the Group 2 questionnaire.

4.1 Sampling and response rates

The fieldwork was conducted by Gallup Europe in the Greater Budapest region of Hungary. Focusing on the capital area offered the advantages of reducing data collection costs, and a single sampling frame, including telephone numbers and addresses, allowing us to hold any error from sampling and coverage consistent across the experimental groups. Of course, restricting our research to a largely urban population may limit the inferences that can be drawn from the findings to the Hungarian population as a whole, and maybe more so if we

wish to generalise to other countries that participate in the ESS. However, while we may expect some cultural differences in the prevalence of certain types of response effect (e.g. Hui and Triandis 1989; Johnson and van der Vijver 2003), we had no reason to assume that the hypothesised mechanisms underlying the mode effects we were interested in would operate differently in this population compared to elsewhere. For this reason, we were content that the study, despite this limitation, would provide useful direction for future related studies in other contexts.

An equal-probability sample of fixed residential phone numbers within the defined area was selected. Each unit was randomly allocated to one of the 3 treatment groups. At each contacted household, one person aged 15 or over was randomly selected for interview using the last birthday method. In total, 515 respondents were interviewed face-to-face using showcards, 518 respondents were interviewed face-to-face without showcards and 685 were interviewed on a fixed-line telephone.[1] The response rate for the telephone group was 32% and that for the two face-to-face groups combined was 33% (AAPOR response rate 1, AAPOR, 2011). The interviews were carried out by 94 interviewers who completed on average 18.3 interviews each. Apart from two interviewers who did both telephone and face-to-face interviews, interviewers worked either on the telephone group (31%), both face-to-face groups (56%), or just one of the face-to-face groups (11%). All analyses presented are adjusted for the clustering of respondents in interviewers.

Face-to-face and telephone surveys may differentially attract respondents with different characteristics (Holbrook, Green and Krosnick 2003). For this reason, a preliminary concern in our analysis was to establish whether there were systematic differences in sample composition across the experimental groups based on the socio-demographic variables at our disposal. The first three columns in Table 1 compare the characteristics of the two face-to-face samples. There were no differences in the characteristics of these two samples, which is to be expected since the only difference between them was in whether or not showcards were used. The latter three columns in Table 1 therefore compare the combined face-to-face samples with the telephone sample. The telephone sample had a significantly lower proportion of men, manual workers and respondents with low education levels. There were

[1] Respondents in the telephone sample were first asked a screener question about whether they owned a mobile telephone. A random half of mobile owners were then called on their mobile for the interview. The other half were interviewed on their landline. Our analyses are based on the landline interviews and we use weights to adjust for the under-representation of mobile phone owners in the landline sample.

however no differences across modes in mean age and the proportion in work. In all subsequent analyses we adjust for differences in sample composition, by using multivariate models and including controls for socio-demographic characteristics.

Table 1: Socio-demographic characteristics by mode

	F2F showcards	F2F no showcards	P-value	F2F combined	Telephone	P-value
Male (%)	38.4	42.5	0.1814	40.5	32.5	0.0063
Mean age (years)	55.8	56.4	0.5534	56.1	55.3	0.4014
Currently in work (%)	50.1	48.5	0.4972	49.3	49.0	0.9133
In manual occupation (%)	35.5	36.7	0.6919	36.1	25.4	0.0024
Less than high school (%)	28.0	26.8	0.6244	27.4	16.7	0.0000

Notes: P-values from Wald tests of the equality of sample means (in each case for the two columns to the left), adjusted for clustering in interviewers and weighted for telephone sampling design.

4.2 Questionnaire

The interviews consisted of a subset of questions from the core questionnaire of the ESS, and covered a variety of topics relating to social attitudes and values, and political involvement. These included measures of social and political trust; political self-efficacy; life satisfaction; trust in institutions; religiosity; attitudes to immigration, gender roles, gay rights and obedience to the law; and behavioural measures including time spent watching television daily; time spent watching news programmes; voting and party voted for; and frequency of religious service attendance. The survey also included socio-demographic measures including gender, age, education, occupation and income.

For group 1, the question and showcard design was essentially identical to the ESS round 2 questionnaires,[2] though the experimental questionnaire was considerably shorter.[3] For groups 2 and 3, the questions using showcards were modified so that they could be administered orally. Different kinds of adaptations were needed and are documented in column 3 of Appendix Table A1. For most questions, the interviewer could simply read out the response categories presented on the showcard, or provide a description of the response scale to be used. For a limited number of more complex items or questions with long lists of response alternatives, the format was changed. This involved either a) breaking the original question down into 'branched' sub-questions (occupation); b) converting the question into an

[2] Round 2 questionnaires and showcards are freely available from the ESS data archive: http://ess.nsd.uib.no/

[3] For the questionnaires used in this experimental study, see the Appendix in Jäckle et al (2006).

open-ended format (time spent watching TV, income); or c) reducing the number of response categories (frequency of church attendance). For all items, interviewers could record a refusal to answer or a 'don't know' response, as is standard practice on the ESS, however, a 'no opinion' option was not offered explicitly to respondents.

Items included in the questionnaire were selected to provide a variety of indicators of data quality across the two modes, based on those used in other mode comparison studies (e.g. de Leeuw 1992; de Leeuw and van der Zouwen 1988). A further criterion was to select items that we believed would be most likely to be susceptible to mode effects, based on either the sensitivity of the question topic or the format of the response scales or categories. We included questions on topics that have been shown in other studies to prompt socially desirable answers, such as income (Groves 1989), religious service attendance (Hadaway, Marler, and Chaves 1993), voting (Cassel 2004; Karp and Brockington 2005; Krosnick 1999); time spent watching television news and interest in politics (Holbrook, Green, and Krosnick 2003); and others we thought would be likely to (e.g. immigration and gender equality). Questions with a variety of response formats were selected, some of which we expected might encourage shortcutting (based on Krosnick 1991) – e.g. lists of nominal categories that might lead to response order effects; batteries of items using the same response scale that might encourage non-differentiation; and attitudes statements with agree-disagree scales that might encourage acquiescent responding.

4.3 Analytic approach and indicators of satisficing and social desirability

Prior analyses showed that for about one-third of items included in the ESS modes experiment, response distributions differed significantly between the telephone and the face-to-face showcard groups (Jäckle et al 2006). To test our hypotheses about the causes of mode differences, we tested for differences in the extent of satisficing and social desirability bias between modes. We examined four indicators of satisficing: acquiescence, non-differentiation, recency and primacy, and two indicators of social desirability bias: selection of desirable responses and selection of undesirable extremes. For each of the six measures, described in more detail below, we constructed an indicator that took values from 0 to 1. The items used to create each of the indicators are summarized in Appendix Table A1.

Acquiescence refers to a tendency to agree with or accept any assertion, regardless of its content (Couch and Keniston 1960). It was identified by Krosnick (1991) as a possible form of weak satisficing, because selecting the agree response may result from respondents finding

it easier to generate reasons to agree with a statement than to generate counter-arguments (though other explanations for acquiescence have been put forward – see e.g. Javeline 1999). Using a battery of six items (labelled from 1 'agree strongly' to 5 'disagree strongly', see Appendix Table A1), we calculated the proportion of items to which the respondent answered 'agree'.

Non-differentiation occurs when respondents rate items in a battery at the same point on the common response scale (Locander, Sudman and Bradburn 1976). Krosnick (1991) identified it as a form of strong satisficing, because respondents who are shortcutting can simply select a response category that seems appropriate for the first item in the battery and stick to that response for all other items, to avoid considering each question separately. We used two batteries of questions (one 7 item battery using an 11 point response scale, and one 6 item battery using a 5 point response scale, see Appendix Table A1) to measure the extent of non-differentiation. For each battery of items we calculated the maximum number of identical ratings made by the respondent and divided it by the number of items in the scale to obtain a variable ranging from 0 to 1. We then created an overall index by averaging the scores from both scales.

Response order effects occur when respondents' answers are influenced by the order in which response options are presented, with the direction of the bias (primacy versus recency) depending on the sensory channel of presentation (visual, as on a showcard, or oral, as in a telephone survey). We measured the extent of 'primacy' effects by calculating the proportion of first-category responses to 12 items where the same response categories were offered for the visual and aural treatment groups. Conversely, we measure the extent of 'recency' as the proportion of last-category responses in the 12 items (see Appendix Table A1).

Social desirability bias. We used two indicators of social desirability bias. Firstly, we created an overall indicator using the proportion of socially desirable responses given to 23 items in the questionnaire that seemed likely to have shared social desirability connotations (see Appendix Table A1 for the items included and the categories classified as desirable). Note that we did not test social desirability connotations empirically, though many of the topics selected have been shown in other studies to invite socially desirable answers. Secondly, we created an alternative indicator based on respondents' willingness to express extreme views. For items where the end-point of a response scale corresponds to an extreme view, selecting the end-point can be seen as an indicator of the willingness to report sensitive information. On this assumption, we calculated the proportion of undesirable first- or last-

category responses given to the 23 items (see Appendix Table A1).

The means and standard deviations for the six indicators in the three experimental groups are documented in Table 2. To test for differences in the extent of satisficing and social desirability bias between modes we used OLS regressions with a mode dummy variable as the main independent variable. Additional variables were included to control for differences in sample composition between modes: dummy for male, age in years, dummy for whether respondent had been in work in the last 7 days, dummy for manual occupation, and dummy for education less than high school. The models were estimated separately for a) the telephone and face-to-face (with showcard) samples, b) the telephone and face-to-face (without showcard) samples, and c) the two face-to-face samples.

Table 2: Summary statistics for indicators of satisficing and social desirability bias

	Telephone		F2F Showcard		F2F No showcard	
Indicator	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
ACQ - Acquiescence	0.190	0.174	0.211	0.205	0.217	0.191
NOND - Non-differentiation	0.445	0.105	0.467	0.122	0.463	0.117
PRIM - Primacy	0.247	0.131	0.237	0.145	0.226	0.150
REC - Recency	0.109	0.110	0.157	0.134	0.147	0.142
SDB - Social desirability bias	0.445	0.143	0.413	0.138	0.411	0.143
UNDEX - Undesirable extremes	0.119	0.092	0.150	0.113	0.140	0.105

5 Results

Table 3 documents the results of testing for differences in the extent of satisficing and social desirability bias between the telephone and face-to-face samples (with showcards). This comparison can be thought of as the overall 'mode system' effect, where differences between modes in the proximity of the interviewer and the question stimulus are confounded. The results suggest that social desirability bias was stronger by telephone: respondents selected a socially desirable response for 2.9% more items than the face-to-face sample did, and conversely selected the undesirable extreme response category for 2.9% fewer items.

The results relating to satisficing are unexpected. First, there are no differences in the extent of acquiescent or primacy responses. Second, the effects for non-differentiation and recency are in the opposite direction from that expected. Compared to face-to-face respondents, telephone respondents selected the same answer category for 2.0% fewer items, and the last response option for 4.7% fewer items.

Are the observed 'mode system' effects large? To get a sense of the magnitude of the mode effects, we briefly examine differences in the extent of satisficing and social

desirability bias between different types of respondents. Consistent with satisficing theory, respondents with low levels of education select the same response option on 2.6% more items in the battery than respondents with higher education. By comparison, the mode effect on non-differentiation is similar in size (2.0%) to the effect of this proxy for 'respondent ability'. As for social desirability bias, men select more desirable responses for 3.8% more items than women, while respondents in manual occupations and with low education select desirable responses for around 6% fewer items than respondents in non-manual occupations or with higher education levels. The mode effect on social desirability bias is therefore similar in size to the difference between men and women, and around half the size of the effects of occupation and education. Given the magnitude of the mode effect it is possible that mode effects could bias research conclusions about differences between socio-demographic groups.

Table 3: Mode system effects

	Telephone vs. F2F showcard					
	(1) ACQ	(2) NOND	(3) PRIM	(4) REC	(5) SDB	(6) UNDEX
Tel (omitted F2F sc)	-0.020	-0.020*	0.011	-0.047***	0.029**	-0.029**
Male	0.022	-0.005	0.020	-0.007	0.038***	-0.008
Age of respondent	-0.001	0.000*	0.002***	0.000	-0.001**	0.001**
Worked in last 7 days	-0.026	-0.010	0.015	0.008	0.010	-0.007
Manual occupation	-0.014	-0.021*	-0.020	0.028*	-0.064***	0.033***
Less than high school	0.030	0.026**	-0.020	0.014	-0.057***	0.028***
Constant	0.245***	0.450***	0.153***	0.128***	0.488***	0.101***
N	1135	1135	1135	1135	1135	1135
r2	0.013	0.029	0.041	0.059	0.150	0.103

Notes: Coefficients from OLS regressions. Dependent variables specified in the text. Standard errors adjusted for clustering in interviewers. Estimates weighted to account for telephone sampling design.
* p<0.05, ** p<0.01, *** p<0.001

Are the mode effects caused by differences in the proximity of the interviewer? If yes, we expect less shortcutting with face-to-face (without showcards) than with telephone interviewing (***H1***), and we expect differences between the two modes in extent of social desirability bias (***H2***).

The results in Table 4 suggest that there are no differences between telephone and the aural face-to-face group in the extent of acquiescent responses or non-differentiation. That is, we find no support for ***H1***.

The differences in the extent of social desirability bias found in the 'mode system' comparison (Table 3) are however replicated. Telephone respondents selected the more desirable response category for 3.5% more items than aural face-to-face respondents, and

selected undesirable extreme categories for 2.1% fewer items. That is, the difference in social desirability bias seems to be largely due to differences in whether the interviewer is physically present or separated by the telephone, supporting ***H2***.

Table 4 in addition documents that there are unexpected differences in the extent of recency effects. Telephone respondents selected the last response option for 3.6% fewer items than face-to-face respondents. This is unexpected, since questions were presented aurally in both groups and we would therefore not expect any differences in response order effects between the two groups.

Table 4: Interviewer effects

	Telephone vs. F2F no showcard					
	(1) ACQ	(2) NOND	(3) PRIM	(4) REC	(5) SDB	(6) UNDEX
Tel (omitted F2F no sc)	-0.028	-0.016	0.026	-0.036*	0.035**	-0.021*
Male	0.019	-0.004	0.017	-0.011	0.042***	-0.013*
Age of respondent	-0.001*	0.001**	0.001***	0.000	-0.001***	0.001**
Worked in last 7 days	-0.023	-0.011	0.008	0.007	0.007	-0.008
Manual occupation	-0.011	-0.007	-0.016	0.025*	-0.058***	0.029***
Less than high school	0.034*	0.022*	-0.029*	0.007	-0.058***	0.023**
Constant	0.278***	0.430***	0.143***	0.118***	0.489***	0.097***
N	1148	1148	1148	1148	1148	1148
r2	0.020	0.032	0.047	0.036	0.149	0.083

Notes: Coefficients from OLS regressions. Dependent variables specified in the text. Standard errors adjusted for clustering in interviewers. Estimates weighted to account for telephone sampling design.
* p<0.05, ** p<0.01, *** p<0.001

Are the mode effects caused by differences in the question stimulus? If yes, we expect differences in the extent of shortcutting with and without showcards (***H3***), and we expect primacy effects with showcards and recency effects without showcards (***H4***).

The results in Table 5 suggest that there were no differences in the extent of acquiescence and non-differentiation between the face-to-face groups with and without showcards. That is, we find no support for ***H3***. We also find no support for ***H4***, in that using a showcard did not produce differences in the extent of primacy and recency effects. Table 5 in addition documents that, as one would expect, the use of showcards did not affect social desirability bias.

As an extension to ***H1***, we further tested whether mode differences were larger among respondents with lower cognitive ability. We used both respondent age (whether aged 65+ vs. younger) and level of education (whether less than high school vs. higher) as proxies for

lower ability. We replicated the models in Tables 3-5 including an interaction between the mode indicator and the ability indicator, running separate models using age and education as proxies for ability. In the 24 models tested[4] the interactions between mode and ability was significant only once. The results (table not shown) therefore did not provide support for the hypothesis that mode effects are stronger among respondents with lower cognitive ability.

Table 5: Showcard effects

	F2F showcard vs. F2F no showcard					
	(1) ACQ	(2) NOND	(3) PRIM	(4) REC	(5) SDB	(6) UNDEX
SC (omitted no sc)	-0.008	0.003	0.016	0.012	0.005	0.009
Male	0.012	-0.000	0.008	-0.007	0.023**	-0.006
Age of respondent	-0.001	0.001***	0.001**	0.000	0.000	0.000
Worked in last 7 days	-0.025	0.005	-0.001	0.017	-0.003	0.003
Manual occupation	-0.003	-0.017	-0.015	0.024*	-0.057***	0.032**
Less than high school	0.014	0.024*	-0.021	0.032*	-0.053***	0.040***
Constant	0.267***	0.401***	0.170***	0.109***	0.456***	0.109***
N	973	973	973	973	973	973
r2	0.005	0.029	0.024	0.029	0.109	0.075

Notes: Coefficients from OLS regressions. Dependent variables specified in the text. Standard errors adjusted for clustering in interviewers. Estimates weighted to account for telephone sampling design. * p<0.05, ** p<0.01, *** p<0.001

6 Summary and discussion

The design of our experimental study enabled us to disentangle the effects of two major differences between face-to-face and telephone interviewing on measurement: differences in the proximity of the interviewer, and differences in the nature of the question stimulus. The results suggested that differences between face-to-face interviewing using showcards (the standard methods used for the ESS), and telephone interviewing were mainly caused by effects of interviewer proximity on respondents' willingness to give socially undesirable responses: telephone respondents were more likely to give answers that were more desirable. Contrary to our expectations and to results from other studies (e.g. Holbrook et al 2003), interviewer proximity had no effect on the extent of satisficing in our study. Similarly unexpectedly, variation in the extent of satisficing by mode was no different between respondents with lower and higher cognitive ability. Showcards did not appear to contribute to the overall differences between face-to-face and telephone interviewing: whether or not a

[4] The 24 models covered all combinations of the two ability indicators (age, education), the four satisficing indicators (acquiescence, non-differentiation, recency, primacy), and the three mode comparisons (Tel-F2F showcard, Tel-F2F no showcard, F2F showcard-F2F no showcard).

showcard was used had no effect on the extent or nature of satisficing. Several limitations of our study warrant discussion and have implications for future research.

Mode differences in the extent of satisficing

First, satisificing theory predicts that respondent motivation to execute the various components of processing will decline as the interview progresses, such that the likelihood of satisficing will increase with questionnaire length, and shortcutting will be more likely on items placed towards the end of the questionnaire (Krosnick 1991). There have been few attempts to test this empirically (although Galesic and Bosnjak (2009) and Yan et al. (2011) compare data quality across questionnaires of different lengths), and nor do we here. Indeed, the questionnaires used in the present study were short by comparison with those in the study by Holbrook et al (2003) (our interviews lasted around 25 minutes on average compared with their one hour interviews) and we did not experimentally vary the position of items in the questionnaire. The design does, however, allow us to test our hypotheses regarding the presence of the interviewer and the nature of the stimulus independent of the potentially confounding influence of interview duration.

Second, as well as being relatively short, the questionnaire in the present study was also quite varied, and the topics and question format changed frequently, which probably made it more stimulating for respondents compared to some longer surveys handling fewer subtopics (the American National Election Study (ANES) questionnaire used by Holbrook et al (2003), for example, is predominantly made up of questions about political attitudes and behaviours). This feature of the design may have helped to maintain respondent motivation over the course of the survey interviews and so limit the urge to satisfice, though of course it limits the generalizability of the findings to the standard-length ESS (which, like the ANES, also takes around one hour to administer). This limitation of the research design highlights the need to investigate empirically the moderating effects of respondent fatigue and declining motivation over the course of the interview on the prevalence of response sets associated with satisficing in different survey modes.

Third, the indicators we used for certain types of response effect may not have provided sufficiently robust tests of our hypotheses. Though other contributors (including Holbrook and her colleagues) have used similar indicators of response quality to investigate the presence of satisficing in surveys, there is considerable variation in how indicators are constructed and a conspicuous lack of agreement about the best approach to the problem. For example, we identified response order effects by comparing the relative likelihood of

selecting a first or last-category response alternative across modes, but without experimental control over the order in which the response options were presented, we were not able to control for question content. A split-ballot design offering the response options in the reverse order for half of the sample in each mode (e.g. as used by Malhotra 2009) would have offered greater control over this potential confound and given us more confidence in our conclusions regarding the relative likelihood of response order effects with different data collection modes. There is also much variation in the methods that have been used to construct indicators of non-differentiation (see, for example, McCarty and Shrum 2000, and Chang and Krosnick 2009), acquiescence, and extreme responses. As indirect indicators of measurement error (such as response effects assumed to occur when respondents satisfice) become increasingly popular in mode comparison studies, the need for greater clarity about the optimal methods for identifying and measuring such effects will become increasingly important for comparing findings across studies. Finally, there are other potential indicators of satisficing that we were not able to construct form our data. For example, "no opinion" responses are also used as indicators of strong satisficing (Krosnick 1991), however only when "don't know" is explicitly offered as one of the response options, which is not the case for the ESS items. In previous work (Jäckle et al 2006) we found no differences in the rates of "don't know" responses between the three experimental groups, although slightly higher "refusal" rates in face-to-face without showcards than telephone (4.7% vs. 4.2%, $P<0.05$).

Fourth, our failure to detect the anticipated satisficing effects in this study may partly be explained by characteristics of the sample. For one, as discussed earlier, the population sampled for the study consisted of residents with listed telephone numbers, living in the Greater Budapest area. These characteristics alone – urban residence (in Hungary), and having a listed telephone number – may have influenced respondents' propensity to satisfice in ways that we were unable to anticipate, and explain why our research findings deviate from those of other studies (though we do not attempt to predict how here). A growing body of research does indeed point to a range of cultural differences in survey response behaviour that may influence respondent preferences for certain response alternatives over others (e.g. Johnson and van der Vijver 2003). Such preferences have often been attributed to cultural differences in social norms surrounding social interactions with strangers (Gordoni, Schmidt and Gordoni 2011), rather than population differences in how the cognitive processes in survey response are executed. Thus, we had no reason to assume based on previous studies that mode characteristics would influence propensity to satisfice differently in Hungary

compared with in the United States, yet we cannot rule out the possibility that other contextual factors distinctive to the survey climate in each country may have played a role (e.g. Holbrook et al's US samples may have been more inclined to satisfice due to greater 'survey fatigue' during general elections, assumed to result from frequent invitations to take part in research). Furthermore, response rates in our study were quite low, so it is possible that those sample members who did take part were more willing to expend the necessary effort to respond to the questions thoughtfully than the non-respondents would have been. Each of these caveats suggest that caution should be taken when generalising our findings relating to satisficing to other populations and to surveys with longer questionnaires.

Question stimulus: effects of showcards

Though it is tempting to feel reassured by the finding that the different questionnaires produced broadly equivalent estimates, we cannot ignore the possibility that we did not observe the expected response order effects associated with visual vs. oral presentation of the response categories due to a failure to effectively manipulate whether and how interviewers used the showcards. Three versions of the questionnaire were delivered to the survey agency responsible for data collection, and – as for the main ESS survey – showcards were provided to go with the face-to-face with showcards version of the questionnaire. No explicit instructions were given to interviewers (beyond the training they received for the study) about *how* the showcards were to be used, and we made no attempt to carry out back checks or validation to ensure that the showcards were in fact used in the way that we had intended. In this respect, the fieldwork protocol matched the standard ESS fieldwork protocol, whereby questionnaires are provided with instructions to use showcards, and the supporting showcards are supplied with the intention that they be used as instructed in the questionnaire. What actually goes on in the field, however, is something of a black box. The questionnaire is scripted in a way that makes it difficult to ignore the presence and purpose of the showcards – and indeed certain questions simply could not easily be answered without the showcard. However, some anecdotal evidence (from briefings with ESS interviewers participating in a pilot study for round 3 of the survey) would suggest that there is indeed considerable variation among interviewers in terms of whether and how they make use of the showcards (perhaps depending on their own preferences, perhaps also in response to perceptions of the respondents' capacity to benefit from their use). While this fact no doubt enhances the ecological validity of our research, the failure to ensure the experimental showcard manipulation operated as intended may well account for our failure to detect the hypothesized

effects associated with using visual aids. Given the widespread use of showcards in personal interviews, and the increasing use of mixed mode data collection, interviewer practice relating to showcards and its impact on the data represents an important area for future research, particularly in efforts to understand interviewer effects on the data.

Mode differences in social desirability bias

While this study failed to detect marked differences in the data between modes resulting from the use of showcards, a consistent finding was that respondents in the face-to-face interviews appeared to give more honest answers to a range of items we assessed as being likely to be susceptible to social desirability bias. Given the robustness of the effect observed in these data, the threat to data comparability from social desirability bias when face-to-face and telephone modes are combined appears to be quite considerable. Understanding the social and cognitive mechanisms underpinning such effects, therefore, represents an important priority for survey methodologists looking for ways to minimise measurement differences between modes. We have discussed several possible explanations that may account for greater prevalence of social desirability bias in telephone compared to face-to-face surveys, but were not able to explicitly test these theories with the data from this study. While the explanations we considered (e.g. greater rapport in face-to-face interviews, more effective communication of survey legitimacy) seem intuitive and may well be sufficient, accounts of the cognitive processes underlying social desirability bias do not fit so well with the data. Notably, motivational accounts of the bias include the idea that respondents edit their true response to survey questions, either to avoid disclosure of personal information to third parties, to bring it in line with social norms, or to deliberately lie to avoid embarrassment (Tourangeau, Rips, and Rasinski 2000; pp. 279-287). A study by Holtgraves (2004), which analysed response times associated with socially desirable responses provides support for this idea. The author concludes that, "when people are more concerned with how their responses might make them look, they do tend to consider their answers more carefully. This does not always affect the particular answer that they give but it does affect how long it takes them respond" (p. 171). If giving socially desirable answers requires respondents to spend longer answering questions, we would expect this to be reflected in increased interview durations among telephone respondents compared with those interviewed face-to-face. In fact, in this study, as in others, this was not the case, suggesting that the respondents had less time available for response editing. While we cannot conclude that the increased prevalence of social desirability bias in the telephone interviews was not the result of conscious response

editing, we cannot rule out the possibility that it may be governed by a more automatic form of processing (e.g. if socially desirable response alternatives are more cognitively accessible, then respondents may find them easier, and so quicker, to select with little consideration). Greater clarity as to the underlying causes of the different effects examined in this study is needed before survey designers can decide the best methods needed to minimise their occurrence.

The social desirability effects observed in these data suggest the threat to data quality on surveys posed by potentially sensitive questions should not be underestimated. We believe preparatory research for surveys considering a switch to a new mode or a mix of modes should be directed at identifying items and question topics that are most likely to be susceptible to the bias and testing empirically their social desirability connotations (as recommended by Holbrook et al. 2003; see also Groves and Kahn 1979). Given the potential for cross-national differences in responses to socially sensitive questions and in the willingness to report potentially sensitive attitudes and behaviours publicly (Gordoni, Schmidt and Gordoni 2011; Johnson and van de Viver 2003), the need to investigate cultural differences in social desirability bias in comparative research is also paramount. At the same time, alternative methods of administrating these items in interviewer-administered surveys (such as with CASI and ACASI) must be explored if interviews are likely to be combined with self-completion methods, which have repeatedly been shown to produce more honest reports on sensitive items (Tourangeau and Smith 1998). Of course, telephone interviews are more restrictive in this respect, a factor to be taken into consideration when deciding whether and how to mix modes.

References

The American Association for Public Opinion Research. (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition. AAPOR.

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. Public Opinion Quarterly, 74, 817-848.

Buelens, B, van der Laan, J., Schouten, B., van den Brakel, J., Burger, J., and Klausch, T. (2012). Disentangling mode-specific selection and measurement bias in social surveys. Statistics Netherlands Discussion Papers, 201211, CBS, The Hague.

Cannell, C., P. Miller, and L. Oksenberg. 1981. Research on Interviewing Techniques. In Sociological Methodology 1981, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Cassel, Carol. A. . (2004). Voting Records and Validated Voting Studies. Public Opinion Quarterly, 68, 102-108.

Couch, A. S. and Keniston, K. (1960). Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable. Journal of Abnormal and Social Psychology 60, 151-174.

Couper, M.P. (2011). The Future of Modes of Data Collection. Public Opinion Quarterly, 75, 889-908

de Leeuw, E. (1992). Data Quality in Mail, Telephone, and Face-to-Face Surveys. Amsterdam: TT Publications.

de Leeuw, E. (2005). To Mix or Not to Mix? Data Collection Modes in Surveys. Journal of Official Statistics, 21, 1-23.

de Leeuw, E., and J. van der Zouwen. (1988). Data Quality in Telephone and Face-to-Face Surveys: A Comparative Analysis. In Telephone Survey Methodology, edited by R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II and J. Waksberg. New York: Wiley.

Dijkstra, W., and Y.P. Ongena. (2002). Question-answer Sequences in Survey-Interviews. Paper presented at International Conference on Questionnaire Development, Evaluation and Testing Methods, Charleston, SC.

Dillman, D.A. (2000). Mail and Internet Surveys: The Tailored Design Method. 2nd ed. New York: John Wiley Co.

Dillman, D.A. (2007). Mail and Internet Surveys: The Tailored Design Method, 2nd ed. Hoboken, NJ: John Wiley Co. (2007 Update).

Duffy, B. (2003). Response Order Effects - How Do People Read? International Journal of Market Research, 45, 457-466.

Galesic, M., and Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. Public Opinion Quarterly, 73, 349-60.

Gordoni, G., Schmidt, P., and Gordoni, Y. (2011). Measurement Invariance across Face-to-Face and Telephone Modes: The Case of Minority-Status Collectivistic-oriented Groups. International Journal of Public Opinion Research Advance Access published June 15, 2011.

Groves, R.M. (1979). Actors and Questions in Telephone and Personal Interview Surveys. Public Opinion Quarterly, 43, 190-205.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley.

Groves, R.M. (1990). Theories and Methods of Telephone Surveys. Annual Review of Sociology, 16, 221-240.

Groves, R.M., and R.L. Kahn. (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic.

Groves, R.M., and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future Public Opinion Quarterly, 74, 849-879.

Hadaway, C.K., Marler, P.L., and Chaves, M. (1993). What the Polls Don't Show: A Closer Look at U.S. Church Attendance. American Sociological Review, 58, 741-752.

Holbrook, A. L., M. C. Green, and J. A. Krosnick. 2003. Telephone vs. Face-to-Face Interviewing of National Probability Samples With Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. Public Opinion Quarterly, 67, 79-125.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. Personality and Social Psychology Bulletin, 30(2), 161- 172.

Hui, C. H., and Triandis, H.C. (1989). Effects of Culture and Response Format on Extreme Response Style. Journal of Cross-Cultural Psychology, 20, 296-309.

Jäckle, A., Roberts, C. and Lynn, P. (2006). Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes (Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project). ISER Working Paper, 2006-41. Colchester: University of Essex.

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

Javeline, D. (1999). Response Effects in Polite Cultures: A Test of Acquiescence in Kasakhstan. Public Opinion Quarterly, 63, 1-28.

Johnson, T. P., and Van de Vijver, F. J. R. (2003). Social Desirability in Cross-cultural Research. In: Harkness, J. A., Van de Vijver, F. J. R., Mohler, P. Ph. (eds.) Cross-cultural Survey Methods. Wiley, Hoboken, NJ

Jordan, L.A., A.C. Marcus, and L.G. Reeder. (1980). Response Styles in Telephone and Household Interviewing: A Field Experiment. Public Opinion Quarterly, 44, 210-222.

Karp, J.A., and D. Brockington. (2005). Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries. The Journal of Politics, 67, 825-840.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of

Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213-236.

Krosnick, J. A. (1999). Survey Research. Annual Review of Psychology, 50, 537-567.

Krosnick, J. A., and D.F. Alwin. (1987). An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement. Public Opinion Quarterly, 51, 201-219.

Locander, William, Seymour Sudman, and Norman Bradburn. (1976). An Investigation of Interview Method, Threat and Response Distortion. Journal of the American Statistical Association 71:269–75.

McCarty, J. A., and Shrum, L. J. (2000). The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures. Public Opinion Quarterly, 64, 271–298.

Miller, P.V. (1984). Alternative Question Forms for Attitude Scale Questions in Telephone Interviews. Public Opinion Quarterly, 48, 766-778.

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

Peytchev, A., Carley-Baxter, L.R., and Black, M.C. (2011). Multiple Sources of Nonobservation Error in Telephone Surveys: Coverage and Nonresponse. Sociological Methods and Research, 40, 138-168

Rogers, T.F. (1976). Interviews by Telephone and In Person: Quality of Responses and Field Performance. Public Opinion Quarterly, 40, 51-65.

Smith, T.W. (1984). A Comparison of Telephone and Personal Interviewing. Chicago: National Opinion Research Center.

Smith, T.W. (1987). The Art of Asking Questions, 1936-1985. Public Opinion Quarterly, 51 (Part 2: Supplement: 50th Anniversary Issue):S95-S108.

Sykes, W., and Collins. M. (1988). Effects of Mode of Interview: Experiments in the UK. In Telephone Survey Methodology, edited by R. M. Groves, P. P. Biemer, L. Lyberg, E., J. T. Massey, W. L. Nicholls II and J. Waksberg. New York: Wiley and Sons, Inc.

Tourangeau, R., and Smith, T. W. (1998). Collecting Sensitive Information with Different Modes of Data Collection. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II & J. M. O'Reilly (Eds.), Computer Assisted Survey Information Collection. New York: John Wiley & Sons, Inc.

Tourangeau, R., L.J. Rips, and K.A. Rasinski. 2000. The Psychology of Survey Response. Cambridge: Cambridge University Press.

Wikman, A., and Waerneryd, B. (1990). Measurement errors in survey questions: Explaining

response variability. Social Indicators Research 22 (199-212).

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. Public Opinion Quarterly, *74, 1027-1045.*

Yan, T., Conrad, F.G., Couper, M. P., and Tourangeau, R. (2011). Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys. *International Journal of Public Opinion Research*, 23, 131-147.

Appendix Table A1: Survey items, response categories, and coding of indicators of satisficing and social desirability bias

Item	Topic	Response format (in aural modes if different)	UNDEX (categories)	SDB (categories)	PRIM	REC	NOND	ACQ
Q1	Time watching TV	0 no time at all – 7 more than 3 hours (open)	7	1-3				
Q2	Time watching TV news	0 no time at all – 7 more than 3 hours (open)	7	1-3				
Q3	Trust people	0 you can't be too careful – 10 most can be trusted	0	7-10				
Q4	Life satisfaction	0 extremely dissatisfied – 10 extremely satisfied	0	7-10				
Q5	Political interest	1 very interested – 4 not at all interested	4	1-2	X	X		
Q6	Political understanding	1 never – 5 frequently	5	1-2	X	X		
Q7	Political opinion	1 very difficult – 5 very easy	1	4-5	X	X		
Q8a	Trust institutions: parliament	0 no trust at all – 10 complete trust					X	
Q8b	Trust institutions: legal system	0 no trust at all – 10 complete trust					X	
Q8c	Trust institutions: police	0 no trust at all – 10 complete trust					X	
Q8d	Trust institutions: politicians	0 no trust at all – 10 complete trust					X	
Q8e	Trust institutions: parties	0 no trust at all – 10 complete trust					X	
Q8f	Trust institutions: EU parliament	0 no trust at all – 10 complete trust					X	
Q8g	Trust institutions: UN	0 no trust at all – 10 complete trust					X	
Q9	Voted last national election	1 yes – 2 no	2	1				
Q11	Immigration: same ethnicity	1 allow many to come and live here – 4 allow none	4	1-2	X	X		
Q12	Immigration: different ethnicity	1 allow many to come and live here – 4 allow none	4	1-2	X	X		
Q13	Immigration: poor outside EU	1 allow many to come and live here – 4 allow none	4	1-2	X	X		
Q14	Immigration: impact on economy	0 bad for the economy – 10 good for the economy	0	7-10				
Q15	Immigration: impact on culture	0 cultural life undermined – 10 cultural life enriched	0	7-10				
Q16	Immigration: impact on living standards	0 worse place to live – 10 better place to live	0	7-10				
Q17a	Gender role: mothers should not work	1 agree strongly – 5 disagree strongly	5	1-2	X	X	X	X
Q17b	Gender role: men responsible for family	1 agree strongly – 5 disagree strongly	5	1-2	X	X	X	X
Q17c	Gender role: men more right to jobs	1 agree strongly – 5 disagree strongly	1	4-5	X	X	X	X
Q17d	Gender role: parents should not divorce	1 agree strongly – 5 disagree strongly	5	1-2	X	X	X	X
Q18a	Homosexuals free to live own lifestyle	1 agree strongly – 5 disagree strongly	5	1-2	X	X	X	X
Q18b	Law should always be obeyed	1 agree strongly – 5 disagree strongly	5	1-2	X	X	X	X
Q19	Religiosity	0 not at all religious – 10 very religious	0	6-10				
Q20	Church attendance	1 every day – 7 never (1 – 4)	4	1				
Q28	Net household income	1 <EUR150 – 12 ≥ EUR10,000/month (open)	1	6-10				

Notes: The numeric items asked as open-ended questions in the aural modes (q1, q2, q28) were coded to correspond to the face-to-face showcard categories. For q20 the 7 showcard categories were collapsed to correspond to 4 the aural categories. See Table 2 for the abbreviations heading columns 4 to 9.