Designing equivalent questionnaires for a mixed mode European Social Survey: Report on the findings of ESS mode experiments

ESSi JRA1, Deliverable 3

Caroline Roberts, Centre for Comparative Social Surveys, City University
April 2008

1. Summary

One of the biggest challenges in contemplating a change in data collection mode in the context of an existing time series is ensuring that there will be no disruption to the continuity of the data. The central coordinating team of the European Social Survey has long recognised this challenge and, since the first round of the survey, has been engaged in a programme of research investigating how different modes could be used on the survey without affecting the quality of the data collected. This report reviews findings from this research relating to how the design of the questionnaire influences the likelihood of differential measurement errors across modes and highlights areas in need of further investigation.

2. Background to this report

This report forms one of the outputs of a collection of related activities funded under the ESSi Joint Research Activity on data collection modes (JRA1) aimed at (1) designing question formats that are relatively insensitive to the effect of mode, and (2) developing guidelines as to the necessary properties of such questions.

In the original proposal for this research, our plan was to investigate the effect of varying the design of questions on the size of measurement errors in different modes, in the context of a mode comparison experiment funded under ESSi. Based on the results of our inquiry, we were then planning to explore the factors underlying any persistent mode effects (i.e. those we could not eliminate despite our efforts to mitigate them through questionnaire design) using qualitative techniques such as cognitive interviewing and behaviour coding. This deliverable was intended to fit between these two stages of the research – reporting on the outcome of the question design experiments and assessing the need for further research.

In the end, based on the findings of our previous research, we decided to change the original proposal for the experiment to be funded under ESSi and instead of focusing on the problem of mode effects (i.e. differential measurement error across modes), we decided to focus on the practical challenges of switching modes (namely the problem of questionnaire length for fielding the ESS by telephone). Because of this switch in focus, no question wording experiments of the kind we had originally envisaged were incorporated into the study.

Despite this modification to our work plan, we now have data from 3 separate ESS studies on data collection modes that can inform our understanding of the problems involved in adapting the standard face-to-face questionnaire for use in alternative modes, and of potential remedies to these problems. In this report, therefore, I review the key findings relating to questionnaire design issues from each of these studies, and derive recommendations for future research aimed at tackling any particularly problematic aspects of questionnaire design identified. I begin by describing what we know about the types of measurement error that would be most likely to affect the ESS questionnaire if it were to be administered in different modes.

3. The design of the ESS questionnaire

Before proceeding, it is worth briefly describing the current design of the ESS questionnaire. The questionnaire consists of two main parts: the 'core' questionnaire, which includes four fixed modules of questions repeated at every round, addressing the following topics: (module A) media and social trust; (module B) political interest, orientation and participation; (module C) subjective well-being, social exclusion, national, ethnic and religious identity; and (module F – placed towards the end of the questionnaires) socio-demographic profile. The second part of the questionnaire consists of 'rotating modules' (typically two at each round), addressing different substantive topics (decided by means of a competition in which cross-national teams of researchers submit proposals to the Scientific Advisory Board), which may be repeated in future rounds of the survey in order to contribute to the time series (to date there have been no repeated modules).

In addition to the six modules that make up the main questionnaire, respondents are also asked to complete a short 'supplementary questionnaire', which is either administered by the interviewer immediately after the main interview, or is left with respondents for self-completion to be collected at a later stage. The supplementary questionnaire contains one fixed module – (module G – consisting of the Schwarz Human Values Scale) and also a short module of test questions (module H) which so far has been used for MTMM experiments designed to evaluate the reliability and validity of different question forms across participating countries (see Saris and Gallhofer, 2007).

With the exception of the supplementary form, the ESS questionnaire was designed to be administered as a face-to-face interview[1] (using either PAPI or CAPI). This means it has a number of specific features that might make it difficult to administer in other modes without some adaptations. These features include: (a) its length – an ESS interview typically lasts around one hour; (b) its complexity – most questions are applicable to all, but some modules do contain routing/ skip patterns (in particular, module F), and some have included randomised split-ballot designs (notably, module H); (c) the length of the questions and lists of response options; (d) its reliance on showcards as visual aids. In this report I consider the implications of some of these questionnaire design features for switching or mixing modes on the ESS. First, I discuss more generally the types of problems associated with mixing data collection modes.

[1] For details about questionnaire development on the ESS, see http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=62&Itemid=96 .

4. Mode effects in surveys

Survey designers considering using a mix of data collection modes need to take into account the fact that each mode (the main ones being face-to-face and telephone interviewing, and paper and Internet-based self-administered questionnaires) has its own unique measurement properties. Put simply, the mode selected affects the quality of the data collected, producing so-called 'mode effects', or different forms of survey error. Where more than one mode is used to collect data from different sample members, the different measurement properties of the modes used are confounded with one another, making it difficult to separate out real differences between groups from differences attributable to the mode of measurement.

As a point of clarification it is helpful to distinguish between three types of mode effect. Mode effects can take the form of (1) coverage error, because not all modes provide access to all members of all populations; (2) nonresponse error, because modes differentially attract different types of respondent; and (3) measurement error, because people respond differently to certain types of survey question when they are administered in different modes. It is the latter type of error that forms the focus of this report.

Based on a wealth of research comparing responses to surveys carried out in different modes, we are now well-informed about the different kinds of measurement errors associated with different modes and the circumstances in which to expect them (see in particular, Groves, 1979; Groves and Kahn, 1979; Schwarz, Strack, Hippler and Bishop, 1991; de Leeuw, 1992; 2005; Dillman, 2000; Tourangeau, Rips and Raskinski, 2000; Holbrook, Green and Krosnick, 2003; and Roberts, 2007). In summary, differences in responses to surveys carried out in different modes come about when characteristics of the mode influence

1. the extent to which respondents feel comfortable to answer openly and honestly to questions that may be of a sensitive or personal nature; and
2. the likelihood of respondents exerting the required cognitive effort to answer the survey questions carefully (see Holbrook et al., 2003; Jäckle, Roberts and Lynn, 2006; Roberts, Jäckle and Lynn, 2006).

These twin influences affect the likelihood of respondents' answers being affected (respectively) by social desirability bias and a range of response strategies associated with 'satisficing' (see Krosnick, 1991) – i.e. shortcutting cognitive processes involved in answering survey questions. Examples of satisficing strategies include acquiescent responding (e.g. always agreeing); non-differentiation, in which respondents rate a battery of items on the same point of a response scale; and selecting the first satisfactory response option, which, depending on the nature of the stimulus (visual or aural), can manifest itself in the form of so-called primacy and recency effects.

To date, our research into the feasibility of mixing modes on the ESS has focused on a relatively modest mixed mode scenario, which would involve allowing certain countries meeting appropriately stringent criteria to switch from face-to-face interviewing to telephone interviewing (a popular request made prior to round 1 fieldwork by certain participating countries where face-to-face data collection is relatively seldom practiced). This has allowed us to concentrate our research on some of the specific issues involved in contemplating a mix of face-to-face and telephone interviewing in a single survey. I focus on these issues in the next section.

5. Combining face-to-face and telephone interviews – effects on measurement

Groves (1979) distinguishes between measurement error arising from 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). This distinction turns out to be particularly helpful when considering the challenges of mixing interviewer-administered modes to collect data from different sample members in a survey[2], because the mode of interviewing – whether in-person or by telephone – has the capacity to influence both actors and questions to produce differences in the quality of the recorded data. In this section I consider how telephone interviewing affects actors and questions differently from face-to-face interviewing and the implications this has for data quality.

Influence of mode on the interviewer-respondent interaction

One of the principal differences between the interview modes is that telephone and face-to-face interactions differ with respect to their 'channel capacity' (Williams, 1977; Groves and Kahn, 1979). Channel capacity refers to the different media through which the interviewer and respondent can communicate. As well as using the audio channel, in-person interviewers can make use of a wide range of visual cues and non-verbal signals not available to telephone interviewers to facilitate communication, help build up rapport, keep respondents motivated and engaged and to slow the pace of the question-answer exchange (Groves, 1989; Holbrook et al., 2003). Each of these characteristics helps to enhance the quality of the data collected in face-to-face interviews, by encouraging respondents to take their time and concentrate on answering the questions. By contrast, telephone interviews can often seem rushed by comparison, and the absence of visual cues and non-verbal feedback can have the effect of increasing the cognitive burden on the respondent and the amount of effort needed to answer the questions conscientiously.

A second important difference between face-to-face and telephone modes of interviewing is the level of intimacy of interactions in each mode (Groves and Kahn, 1979). This refers in particular to the quality of the relationship established between the interviewer and the respondent over the course of the interview. Telephone interviews tend to be more impersonal than those conducted face-to-face (Groves 1989) and there are fewer opportunities for establishing rapport. This is partly explained by the social distance between the actors (de Leeuw and van der Zouwen 1988) and the different communication channels available, but other factors are likely to be important too (e.g. Groves (1990) discusses differences in the social norms surrounding interactions with strangers in person compared with over the telephone).

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. Contrary to the assumption that the presence of the interviewer reduces the privacy of the reporting situation for

[2] Mixing modes does not generally present a problem where all respondents are asked the same questions in the same mode (even if different modes are used to administer different parts of the questionnaire) – see de Leeuw, 2005; Dillman, 2000.

respondents and increases socially-desirable reporting[3] (Tourangeau and Smith 1996; 1998), it turns out that the better rapport established in face-to-face interviews and the greater opportunities it provides for reassuring respondents of the legitimacy of the survey and confidentiality of the data, make it a more effective method than telephone interviewing for obtaining potentially sensitive information from respondents (e.g.de Leeuw and van der Zouwen 1988; Groves and Kahn 1979; Holbrook, Green, and Krosnick 2003). A growing number of mixed mode studies comparing data from face-to-face and telephone interviews are finding more evidence of social desirability bias among telephone respondents, despite the fact that the relative remoteness of the interviewer might give the impression of greater privacy (see e.g. Smith, 1984; De Leeuw and van der Zouwen, 1988; Holbrook et al., 2003; Jäckle, Roberts and Lynn, 2006).

Influence of mode on how questions are asked

As well as influencing the nature of the interaction between the interviewer and respondent, telephone interviewing differs from face-to-face interviewing in terms of how the questionnaire can be administered (Groves 1979). In particular, whereas in face-to-face interviews it is possible to make use of a variety of visual aids (e.g. 'showcards' or the interviewer's laptop or handheld computer) to ensure the respondent has understood the question and can remember the response options available, in telephone interviews, where communication is restricted to the auditory channel, the interviewer must rely on respondents' ability to recall the information that has been read to them while they formulate an answer. This increases the difficulty of the response task considerably, making it necessary for question stems and lists of response categories to be shorter so that the respondent can hold them in working memory during the answering process. It is not uncommon, therefore, for questionnaires designed for telephone interviews to contain more 'unfolding' or 'branching' type questions (Groves 1990; Krosnick and Berent 1993) that essentially break typical face-to-face survey items down into two or more parts with smaller sets of response options, making them easier to administer aurally. By contrast, the possibility of making use of showcards (cards displaying the available response options) in face-to-face interviews means that even long and complex questions can be relatively easily administered. Many surveys conducted in face-to-face mode have exploited this possibility with a view to improving the accuracy of measurement and the ESS is no exception.

A brief review of the literature, however, calls into question the advantages of using showcards in face-to-face interviews (Miller 1984). There is evidence that interviewers find them useful (Rogers 1976), probably because they help to speed up the response process by minimising the need to repeat the list of options (Jordan, Marcus, and Reeder 1980; Duffy 2003); but the evidence that they facilitate the response process for the respondent is limited. Instead, a number of studies have suggested that using showcards may in fact *increase* the cognitive burden on respondents, who often have to read and absorb the information presented on the card

[3] This conclusion would be consistent with the findings of a wealth of studies that have shown that respondents invited to complete self-administered questionnaires (either on paper or on a computer) are less likely to over-report behaviours or attitudes considered to be socially desirable or to underreport behaviours or attitudes that are considered socially undesirable (Aquilino 1994; DeMaio 1984; Epstein, Ripley Barker, and Kroutil 2001; Jobe et al. 1997; Tourangeau and Smith 1996; 1998).

in a relatively short space of time and may feel pressure to do so quickly, so as not to keep the interviewer waiting (Sykes and Collins 1988; Duffy 2003). Listening to the interviewer may distract respondents whilst reading, or alternatively, reading the showcards may distract respondents from listening to the question (Dijkstra and Ongena 2002). Where many showcards are used in an interview, respondents may find it tiring or confusing to use them (particularly if showcard questions are mixed with those that do not require them), making them more likely to satisfice.

There is also compelling evidence that the visual layout of information on the showcard may bias response selection. One form of satisficing associated with visual presentation of response alternatives is the effect of primacy - the tendency for respondents to select items near the start of a list of alternatives in preference for later items (Krosnick and Alwin 1987). These so-called response order effects arise precisely because of the burden on short-term memory. As respondents read down the list (either on the showcard, or equally, in a self-administered questionnaire), early items are processed more carefully than later items and are consequently more likely to be selected[4]. By contrast, the opposite effect is often observed in data from telephone interviews, with respondents showing a preference for items towards the end of the list of options (referred to as a recency effect) because these are more likely to be retained in the respondent's short-term memory. Groves and Kahn (1979) also found respondents in their face-to-face interviews showed greater preference for scale points that were labelled on the card, compared to telephone respondents for whom the scale was described out loud (Groves 1990).

The increased cognitive burden of telephone interviews associated with the absence of visual cues (either in the form of showcards or positive, non-verbal feedback from the interviewer) imposes further restrictions on data collection in this mode – notably, in terms of *how many* questions can be asked. Telephone interviews are typically much shorter than face-to-face interviews, with survey houses often imposing formal restrictions on questionnaire length. While the primary motivation behind such restrictions is likely to be to increase participation and to avoid break-offs mid-interview, it probably also helps to ensure the quality of the data. In any kind of long survey interview, we would expect some decrease in motivation over the course of the interview, while burden is likely to increase as the interview progresses and the respondent tires (Krosnick, 1991; Holbrook et al., 2003). These effects are especially likely to occur in surveys conducted by telephone.

6. Implications for mixing modes on the ESS

To summarise the above description of how different characteristics of modes can influence the quality of data collected in a survey:

- Telephone interviews are more likely to elicit socially desirable responses compared with face-to-face interviews (though both are more susceptible to the bias than are self-completion modes).

[4] Duffy (2003) presents evidence to suggest that some respondents may develop alternative reading strategies over the course of the interview, for example, where they note more 'popular' response options are displayed near the bottom of the showcard in reversed-order lists.

- Telephone interviews are more likely to elicit satisficing effects, especially where the questionnaire is long.
- Data from face-to-face interviews are likely to be susceptible to effects associated with the use of showcards, such as primacy effects.

If the ESS were to permit countries either to switch to telephone interviewing altogether, or to incorporate telephone interviews alongside face-to-face interviews as part of a mixed mode strategy, data collected in each of the different modes would be likely to be affected by these different kinds of measurement errors, making them incomparable. Our research has attempted to establish the extent of differences in measurement when the ESS questionnaire is administered in different modes and the likely causes of differential measurement errors, with a view to making recommendations about how best to try to reduce them.

As stated earlier, these different types of measurement error come about because different characteristics of the mode affect the amount of effort required to answer the questions accurately and the likelihood that the true response to the question will be reported. Dillman (2000) has argued, however, that differences observed in mixed mode data are often not caused by mode effects per se, but by differences in how questions are constructed in the two types of questionnaire. In comparisons of face-to-face and telephone data, this is evidenced in the fact that showcard questions are among those most likely to exhibit mode effects (Groves 1990; Groves and Kahn 1979), but even relatively small differences in the wording of questions in each mode may lead to differences in response (irrespective of the effect of mode per se). For this reason, in contemplating a switch in mode, we have paid particular attention not only to the possibility of effects caused by characteristics of the mode, but also to the equivalence of the questionnaire to be used in each mode. Our ultimate goal has been to develop questionnaires for use in different modes that produce data that are as comparable as possible, without making changes to the existing face-to-face survey instrument (so as to ensure the continuity of the time series). It is in this context that the activities that form task 2 of JRA1 were proposed, and it is this challenge, which forms the focus of the present report.

7. The ESS mode experiments

JRA1 was designed to build on an existing programme of research on the ESS, already exploring the feasibility of mixing modes in future rounds of the survey. This research was partly funded by the methodological budget from rounds 1 and 2 of the ESS, and partly by the survey organisation, Gallup Europe, who had a shared interest in problems associated with mixing modes in comparative surveys. The ESS-Gallup Mixed Mode Methodology Project consisted of two main phases of experimental work, which have since been supplemented by a further study funded under ESSi. The following provides a brief summary of each phase and the rationale behind the design of each.

a. Phase 1 – Pilot work to investigate the sensitivity of ESS questions to data collection mode (Hungary, 2003)

Description: A mixed mode experiment carried out as a ‘Hall Test’ by Gallup Europe in Hungary (2003), in which a quota sample of participants recruited in the street were randomly allocated to complete a survey questionnaire in two of four modes (face-to-face interview, telephone interview, paper-and-pencil self-completion and web self-completion). After completing the first interview/questionnaire (wave 1), the participants were re-assigned (at random) to complete a second interview/ questionnaire about twenty minutes after the first in a different mode (wave 2). All data collection was carried out on site (four separate hall test events were held, two in Budapest, the others in Györ and Pécs) with the exception of one-half of the wave 2 interviews in telephone and self-completion mode, which were carried out in participants’ homes two weeks after the first event.

Research aims: The primary aim of the study was to investigate mode effects on measurement by comparing responses to questions between different pairs of modes. The within-subjects experimental design made it possible to examine the effect of mode while controlling for the effect of selection bias. The questionnaire included items from the ESS and from the Eurobarometer, chosen on the basis that they seemed particularly likely to be susceptible to certain types of mode effects. The original face-to-face questions were adapted to make them suitable for administration in the other modes and, therefore, to mitigate the most likely mode effects. A ‘built-in’ question-wording experiment was included to test alternative versions of some of the questions, to see which formats were associated with the smallest mode differences.

Methodology: The full experimental design is presented in table 1, along with the achieved number of interviews in each group. The design was not fully crossed – there were no wave 2 web interviews. In addition to the question wording experiment (administered using the two questionnaire versions), a further feature of the design was the in-hall/at-home split for the wave 2 telephone and self-administered interviews described already. The purpose of this feature was to test the effect of carrying out the wave 2 data collection after two weeks, compared with just twenty minutes, as well as the effect of enhancing the ecological validity of the data (though unfortunately these two factors were confounded in the experimental design).

Table 1: Phase 1 experimental design: achieved interviews by mode, questionnaire version and location

			Mode					
Wave	**Location**	**Version**	**Face to face**	**Telephone**	**Self-completion**	**Web**	**Totals**	
1	**In Hall**	**A**	682	616	474	185	**1957**	**1957**
2	**In Hall**	**A**	210	110	120	-	440	
		B	213	115	151	-	479	
		Total	**423**	**225**	**271**	**-**	**919**	
2	**At Home**	**A**	-	173	221	-	394	
		B	-	169	289	-	458	
		Total	**-**	**342**	**510**	-	**852**	**1771**

The question wording experiment deserves some attention here. Four measures were included in the experiment: two attitude items, each consisting of a statement and a 5-point agree-disagree response scale (a) ‘Gay men and lesbians should be free to live their own life as they wish’; b) ‘Whatever the circumstances, the law should always be obeyed’); one behavioural measure of frequency of attendance at religious services; and a measure of net household income. For the attitude measures, version A of the questionnaire had a fully-labelled scale (displayed vertically on the showcard/ SAQs), and version B had labels only for the end points of the scale (and was displayed vertically). For the religious service attendance measure, version A used the standard face-to-face format which includes 7 categories; version B used a shorter list of options, which collapsed the 7 categories into 4. For the income measure, combinations of any of four items were compared across the different questionnaire versions (the combinations varied by mode – see appendix for details).

Key findings[5]: A typology of question types differentiating between (1) sensitive questions, (2) questions with a visual stimulus in some modes and aural stimulus in others, (3) questions with hidden pre-codes, and (4) long or complex questions guided the analysis of data carried out by the project team (see Bryson, O’Shea and Nicolaas, 2003). Based on this typology, the team focused on the following types of mode effect: social desirability bias, primacy and recency, digit preference (defined here as the tendency to select 0, 5 and 10 on an 11-point scale) and item non-response. The results of their analysis were mixed. In most cases, the design of the questionnaires was found to have mitigated many of the anticipated effects, but in other cases, these effects persisted. In still other cases, mode effects that were not expected were observed, leading the project team to conclude that further research was needed to better understand the nature and causes of the effects found. Analysis carried out on behalf of Gallup Europe by Peytcheva and colleagues (2004) further evidenced this mixed pattern of results, but focused on pair-wise comparisons between each of the modes, revealing that data from face-to-face interviews and self-administered questionnaires were more similar, while data from telephone interviews were most different from other modes. These findings were further supported by Kaminska and colleagues’ (2007) analysis.

Analysis of the question-wording experiments also revealed mixed findings. For example, for the first of the two attitude statements, there were no significant differences between the two versions of the response scale (although there was some evidence that in comparisons between responses in telephone and self-completion modes, the fully-labelled scale produced smaller errors). For the second attitude measure, however, it was the scale with only the end-points labelled that produced more comparable data (but then only in comparisons between responses in face-to-face and telephone interviews, and in face-to-face and self-completion modes). No differences were observed between response distributions on the religious service attendance measures (when the 7-category version was recoded to match the 4-

[5] The data were analysed initially by the ESS project team at the National Centre for Social Research (see Bryson, O’Shea, and Nicolaas, 2003 for a full report), and by, and on behalf of, Gallup Europe (see Peytcheva, Manchin, Tortora and Groves, 2004). Subsequently, the data have been analysed by scholars at the University of Michigan and the University of Nebraska (see e.g. Kaminska, Bautista and Serrano, 2007; Serrano, Kaminska, McCutcheon and Manchin, 2007). The analysis of the question-wording experiments presented here was carried out by the author (in consultation with Peter Lynn).

category version) and the pattern of findings relating to the comparisons between different methods of measuring income were also mixed. Few differences were observed, and those that were evident were difficult to interpret because for within-subject comparisons, the mode of data collection was confounded with questionnaire version (i.e. there was no control group re-interviewed in the same mode), as well with the timing of the wave 2 data collection.

Conclusions: The findings of the study suggested that the two modes that differed most from each other were face-to-face and telephone interviewing, while fewer differences were observed in comparisons of face-to-face interviewing with the self-completion modes. A possible interpretation of this finding was that the use of showcards, which provided respondents interviewed in person with a visual question stimulus, made the face-to-face mode more similar to the self-completion modes. This raised concern among the ESS team, given that telephone interviewing seemed the most likely alternative to face-to-face, were the survey to allow data collection in multiple modes. Partly for this reason, the decision was taken to focus on these two modes in the second phase of the research and to postpone the consideration of self-completion modes until later.

A further motivation for the decision to focus on just two modes in phase 2 was the complexity of the experimental design used in phase 1. As noted above, a number of confounds in the design further complicated matters, making it difficult to draw robust conclusions about the nature and cause of the mode differences observed. Given that one of the ultimate aims of the research programme was to try to find ways to mitigate differences in measurement between modes, it was decided that the phase 2 study should have a specific focus on trying to identify the causes of observed differences, so we could develop ways to try to tackle them.

b. Phase 2 – Survey experiment to investigate mode differences between face-to-face and telephone interviewing and their underlying causes (Hungary & Portugal, 2005)

Description: A survey experiment carried out by Gallup Europe in Hungary and Portugal[6] in 2005 (see Jäckle, Roberts and Lynn, 2006 for the full report). The design of the study included three treatments: (1) a face-to-face interview using a reduced-length version of the ESS questionnaire; (2) a telephone interview using the same questionnaire, but adapted for aural administration (i.e. it dispensed with the showcards); and (3) a face-to-face interview using the telephone questionnaire. A random sample of respondents selected from frames containing addresses and telephone numbers covering the Budapest and Lisbon regions were randomly assigned to one of the three conditions[7]. To enhance the ecological validity of the study compared with phase 1, all interviews were carried out in respondents' homes.

Aims: Building on the phase 1 research, the principal aim of the study was again to explore the sensitivity of ESS questions to mode of administration, this time focusing

[6] Data collection in Portugal was carried out by Consulmark, who are partners with Gallup Europe in the Flash Eurobarometer consortium.

[7] The telephone group was further split to allow around one third of respondents to be interviewed on their mobile telephone.

on the comparison between face-to-face and telephone interviewing. A further aim was to try to disentangle the most likely causes of inter-mode differences. The design of the study was such that it enabled us to identify two main types of effect: the effect of differences in the design of the questionnaire, including the use of showcards, and the effect of other characteristics of the mode, the most obvious being the presence/absence of the interviewer. The questionnaire included items selected on the basis that they would be particularly susceptible to measurement errors of different kinds, allowing us to identify a range of effects on data quality. In particular, we included items that would serve as indicators of social desirability bias and various forms of respondent satisficing (including e.g. completeness of responses to open-questions, non-differentiation, acquiescence, and response order effects).

Research design: Table 2 shows the experimental design and the issued and achieved sample sizes for the experiment conducted in Hungary. Note that comparisons between groups 1 and 2 allowed us to identify so-called 'stimulus/ showcard effects', while controlling for other effects of mode. Comparisons between groups 2 and 3 allowed us to identify the effect of mode (particularly the presence/absence of the interviewer), net of the effect of differences in the design of the questionnaire. Comparisons between groups 1 and 3 revealed the overall 'system' effect (i.e. the effect of a telephone interview compared to the standard ESS face-to-face interview), but note that it confounds the differences in how the questions were constructed with differences in the mode of administration. Note also that the design did not allow us to control for differential nonresponse across modes, so any compositional differences between the achieved samples in each mode had to be controlled for at the analysis stage.

Table 2: Phase 2 experimental design: issued and achieved sample sizes

Group	Description	Issued Sample sizes	Completed interviews
	Face-to-face interviews	**3300**	**1033**
1	- with showcards		515
2	- without showcards		518
	Telephone interviews	**2850**	**887**
3	- fixed-line		685
	- mobile		202
		6150	**1920**

Key findings: Our analysis of the data had two goals. The first was to simply assess the overall extent of mode effects in the data, by comparing means and distributions of responses to closed questions. Because of the need to control for socio-demographic differences among the telephone and face-to-face respondents that might otherwise account for any observed differences in responses, we were restricted in terms of the methods of analysis available to us (see Jäckle, Roberts and Lynn, 2008 for further details). We used OLS regression models to examine the effect of mode on mean responses in each of our three comparison groups and Proportional Odds Models (O'Connell, 2006) to examine the effect of mode on response distributions. Overall, we concluded that mode had affected responses to only around 40% of the 28 questions we analysed, though there was evidence to suggest that these two methods had identified different types of mode effect (see Jäckle, Roberts and Lynn, 2006; 2008).

The second goal of the analysis was test hypotheses about the causes of the observed differences (see Roberts, Jäckle and Lynn, 2006), by testing for differences in the extent of satisficing and social desirability bias due to differences in the stimulus (visual vs. aural presentation of response options) and the presence vs. absence of the interviewer. Most of the differences we observed appeared to be attributable to differences in mode (e.g. interviewer presence/ absence) rather than differences in the nature of the question stimulus, which manifested itself as a tendency for increased social desirability bias in the telephone mode. We found little evidence of differences between the modes in the extent of respondent satisficing.

Conclusions: Based on the findings of our analysis, we concluded that the extent and nature of measurement differences between the modes were not as serious as we had first anticipated. Only a relatively small proportion of the items included in the study showed mode effects (and it is noteworthy that these were questions we deemed most likely to be susceptible to mode effects), and the effects we observed were relatively small. Moreover, the mode differences we found on individual items did not affect the results of a number of multivariate analyses we looked at, suggesting that analysts working with the mixed mode data would be unlikely to reach different conclusions to those working with the standard face-to-face data. The mode effects that we did observe appeared to be consistent with previous findings in the literature – namely, an increased propensity for socially desirable responding in telephone interviews compared to face-to-face interviews, suggesting that telephone respondents had not felt sufficiently at ease to disclose their true responses to the questions. However, it was not clear to us whether these effects could be attributed to respondents editing their true responses in light of social desirability concerns, or whether perhaps this response strategy was another form of shortcutting (Jäckle, Roberts and Lynn, 2006; Roberts, 2007).

A number of features of the design of the study limited the strength of the conclusions we could draw from our findings. In particular, the questionnaire we used for the study was substantially shorter than the standard ESS questionnaire (interviews lasted, on average, 15 minutes in face-to-face mode, and just 12 minutes by telephone). This provides the most likely explanation for the absence of satisficing effects observed in our data. Similarly, our conclusions about the effects of showcards on response are limited to the extent that we do not know enough about how interviewers use showcards in the field, and whether in fact they were used as intended in this particular study. A further difficulty stemmed from the problem of selection bias. Differential nonresponse across modes (an effect we would expect to see in mixed mode surveys, but which can be avoided in mode comparison studies by using experimental designs with repeated measures) meant that in order to detect mode effects in our data, we needed to control for the composition of samples in our analyses. The different statistical methods we used, however, led to different conclusions about the extent and nature of mode effects in the data (Jäckle, Roberts and Lynn, 2008). This highlights the inherent difficulties of conducting mode comparisons, and consequently, the problem of drawing generalisable conclusions from a single study.

c. Phase 3 – Survey experiment to investigate the practical challenges of conducting the ESS by telephone (Cyprus, Germany, Hungary, Poland and Switzerland, 2006/2007)

Description: Having investigated the problem of mode effects in data collected by face-to-face and telephone interviewing, we turned our attention to the practical challenges involved in switching modes on the ESS. Our phase 3 research consisted of a survey experiment in five countries (Cyprus, Germany, Hungary, Poland and Switzerland) carried out during the ESS round 3 fieldwork period (2006/2007), designed to test the feasibility of conducting the ESS by telephone. The specific focus of the research was on the length of the interview, and whether or not this would pose a barrier to successful telephone administration of the survey. If it proved impractical to field the full ESS interview in a single telephone interview, then two alternative approaches might offer a solution: one option would be to split the sample so that half the respondents answered just one of the two rotating modules, and the other half answered the other; another would be to split the questionnaire into two or more parts, and to ask all respondents to participate in two (or more) interviews. We field tested both solutions in this study.

Aims: We wanted to examine a number of practical issues involved in collecting ESS data by telephone. This included assessing the feasibility of conducting a long survey interview; the suitability of potential alternatives to the standard questionnaire design; issues involved in the adaptation of questionnaires to make them suitable for telephone administration; and how best to translate the existing ESS guidelines for fieldwork procedures (regarding e.g. the minimum number of contact attempts, timing of contact attempts, recording the outcome of contact attempts and so on) into equivalent recommendations for conducting the survey by telephone. In this context, our principal objective was to assess the impact a switch to telephone would have on response rates, and to measure the effect on participation of varying participants' expectations about the likely length of the interview (and the actual length of the interview itself). A secondary aim was to assess the quality of the data collected; given our hypothesis that long interviews by telephone would be more likely than the short interviews we conducted in the Phase 2 study to induce satisficing among respondents. Given this focus, we were not specifically interested in measuring mode effects in terms of differential measurement error between modes, but we were interested in looking at some of the specific measurement properties of telephone interviews. We did not consider experimenting with different versions of specific survey questions as we had no firm basis for comparison with the face-to-face data, but the process of adapting the face-to-face questionnaire for the telephone highlighted questions that might be particularly likely to give rise to errors in the data (namely, complex showcard questions), so a further aim was to evaluate the success of our 'translations'.

Research design: Probability samples yielding approximately 1000 eligible cases[8] in each of the five countries were randomly assigned to one of three treatment groups. The three groups varied in terms of the length and structure of the questionnaire, and sample members were informed in the survey introduction how long the interview

[8] The sample design varied across countries depending on the availability of suitable frames and funding. National teams were instructed to select a starting sample of 1000 eligible cases. Depending on the sample design, issued sample sizes varied cross-nationally. Details are shown in table 3.

would be expected to last (they were then invited either to make an appointment to participate in the interview at a time convenient to them, or to start the interview straight away). Group A received the standard full-length ESS interview, which lasts about 60 minutes. Group B received the full interview, minus one of the rotating modules, resulting in an interview expected to last around 45 minutes. Group C received the full interview, split into two parts expected to last around 30 minutes each. At the survey introduction, group C respondents were told the interview would last 30 minutes. Only at the end of the first interview were they asked if they would be willing to participate in a second 30 minute interview (either straight away or on a separate occasion). Table 3 shows the issued sample sizes for each group in each of the participating countries.

Table 3: Phase 3 issued sample sizes, dispositions at last call and response rates by country and questionnaire version

CALL OUTCOME	**Cyprus**	**Hungary**	**Germany**	**Poland**	**Switzerland**
Total issued sample	1056	1000	1545	1422	859
Most recent disposition:					
Contact, interview	83	252	369	339	342
Non-contact	388	101	250	270	62
Contact – no interview	34	78	41	140	174
Refusal	318	569	860	211	247
Not eligible	59	-	25	462	34
Number of complete interviews	65	210	329	292	293
Response rates:					
Version A – 60 mins (%)	8.5	18.0	20.3	32.1	37.9
Version B – 45 mins (%)	5.3	22.0	25.0	37.0	39.5
Version C (both parts) (%)	5.2	23.5	21.3	25.6	26.8
Version C (all part A) (%)	8.8	31.5	25.2	32.4	50.2
Overall ESS response rate (complete interviews only) (%)	**6.5**	**21.0**	**21.6**	**30.4**	**35.5**

With the exception of the module removed from the group B questionnaire, all respondents were asked the same questions, which were the same as those in the Round 3 face-to-face questionnaire. In group C, in order to ensure sufficient information was obtained from respondents during the first interview to be able to make some use of the data should they not wish to take part in the second interview, the order of questions was changed such that part of the socio-demographic questions in module F were moved to the first half of the questionnaire. In order to make the questionnaire suitable for telephone administration, almost all questions needed to be adapted to make them suitable for telephone administration. Modifications to the face-to-face questions took the following forms:

1) For all affected questions, references to showcards were deleted, new instructions were added for interviewers to 'read out' response options, and for some (e.g. items with 11-point response scales), a description of the response scale to be used was added.
2) For a small number of questions with many and/or complex response categories, which rely on more elaborate showcards in face-to-face mode, the following types of changes were necessary:

a. converting the question to an open-ended format
b. breaking the question into two parts
c. collapsing certain categories of response to reduce the overall number of options to be read out by the interviewer.

This particularly affected certain socio-demographic question, including main activity, parents' occupation, income and marital status.

Key findings: The data analysis we have conducted to date has been focused on our principal research questions concerning the effect of interview length on response rates and on the quality of the data collected by telephone (see Eva, Roberts and Lynn, and Roberts, Eva, Allum and Lynn, both forthcoming). Our preliminary findings in relation to response rates suggest that as expected interview length is reduced, willingness to participate increases. However, the pattern of results across countries was mixed, which suggests our manipulation of expectations of interview length may not have been as effective as we hoped. Table 3, which shows the final disposition of samples (based on the outcome of the last call[9]) summarises these initial results. We are also looking at the effect of interview duration on response quality in a module of questions on personal and social wellbeing. The next step will be to examine the effect of adapting the questionnaire on response distributions, where comparisons with the face-to-face data are possible.

Conclusions: As our analysis of these data is still ongoing, it is not possible to draw any firm conclusions from the study at this stage – particularly in relation to questionnaire design issues. However, the initial results lend support to the argument for a reduction in questionnaire length if the ESS were to use telephone data collection in its future rounds. Sample members invited to participate in shorter interviews were more likely to agree to participate, but splitting the questionnaire into two parts did not prove to be a successful means of administering the entire ESS interview. Assessing the extent to which interview duration mediated the likelihood of certain types of measurement errors is an immediate priority. Only then can we substantiate our conclusions from the phase 2 research regarding the extent of respondent satisficing in the ESS.

In addition, there is a pressing need to investigate the effect of our adaptations to the questionnaire on the data collected, particularly for those questions that had to be changed substantially due to the design of their accompanying showcards. In almost all cases, these problematic questions were ones that aim to measure respondent's socio-demographic characteristics, making the need to validate their equivalence with their face-to-face versions all the more urgent. Analysis of mixed mode data relies on the possibility of controlling for compositional differences in the samples responding in each mode. This possibility is reduced when the questions used to measure important characteristics are not strictly comparable. For this reason, at the next stage of our analysis, we will turn our attention to assessing the 'success' of our adaptations to complex showcard questions, and to deciding what further development work will be needed to ensure our new measures are as comparable as possible as those used in the face-to-face survey.

[9] Our ongoing analysis is using call record data to compute final dispositions. In practice, outcome of last call has been shown to be a good proxy – McCarty (2003) for example, demonstrated that response rates based on most recent call outcome differed from final disposition based on call record analysis by just 2-5%.

This work is complicated by the fact that our opportunities for making comparisons between the telephone data and data from the round 3 face-to-face ESS are restricted because the effect of our questionnaire adaptation will be confounded with other characteristics of the data collection modes in our two data sources. In other words, unlike in our phase 2 experiment, we have no face-to-face respondents interviewed using our telephone questionnaires (i.e. without showcards) to allow us to disentangle the effects of the questionnaire from other possible influences on respondents' answers. We hope that a preliminary analysis of our data will alert us to the existence of likely problems with the items in question, but we recognise the need for further research to address this issue more thoroughly and systematically.

8. Directions for future research

The original purpose of this paper was to report on the findings of question wording experiments from the ESSi JRA1 mixed mode research (task 2), with a view to providing recommendations for the next phase of the research. Because we did not, in the end, carry out any such experiments in the most recent phase of our research, preferring instead to focus on some of the more practical challenges involved in switching to telephone interviewing, we have not been able to uniquely focus on data of the kind we had originally envisaged for this deliverable (i.e. data that allows us to compare the size of measurement errors across modes arising from questions presented in different formats). Instead, we have drawn on a wider range of evidence from our previous studies to inform our understanding of mode effects on measurement, and to highlight problems with administering the ESS questionnaire in different modes (in particular, by telephone) that we believe would benefit from further investigation. In this section, we summarise the areas requiring particular attention.

One of the main lessons to come out of the phase 1 research was the need to understand better the causes of mode differences in our data. Phase 2 was designed to address this need, but a number of questions remain unanswered:

1) ***What effect, if any, do showcards have on data collected in face-to-face interviews?***

The results of our phase 1 research suggested that showcards might help to enhance the equivalence of data collected by face-to-face interviews and self-completion modes because of the common visual stimuli each provides. By contrast, the absence of a visual stimulus in telephone mode appeared to make responses in that mode differ most from the other modes. Our attempt to investigate this further in our phase 2 experiment, however, found that the use of showcards had only a minimal effect on the comparability of data from face-to-face and telephone interviews. The discrepancy in these findings raises questions about how showcards are actually used by interviewers and respondents during an interview and the effect of using showcards on data quality. In particular, we cannot be certain that in the showcard condition in our experiment all interviewers used the showcards the way we had intended, and some respondents may have preferred not to use them at all. Because of this, we had little experimental control over the showcard treatment in our phase 2 study and this weakens the robustness of our conclusions about the effect of using showcards on measurement error.

We cannot conclude on the basis of the research we have conducted to date (and the topic has received comparatively little attention elsewhere in the literature) that the use of showcards on the ESS would not give rise to mode effects were telephone interviewing to be introduced alongside face-to-face. Furthermore, our experiences of adapting the ESS questionnaire for the telephone in phase 3 suggest we should be particularly concerned about the design of questions that rely on elaborate showcards for their administration in face-to-face mode, as these are the ones that would require the most substantial modifications, making it difficult to ensure their equivalence with the original questions. The fact that items of this kind are often intended to measure key background variables makes it especially important to assess the impact of the two forms of stimuli on data comparability.

Three types of investigation into the effect of using showcards would appear to be particularly beneficial in the development of equivalent face-to-face and telephone questionnaires:

i) An assessment of how showcards are used by interviewers in the field – whether they are used at all, whether they are generally seen as a help or hindrance by interviewers and respondents[10];
ii) Cognitive interviews to assess whether showcards facilitate or unnecessarily complicate the response process for respondents
iii) Experiments to compare questions using visual stimuli with aural equivalents. These could focus on specific questions (e.g. the problematic demographic measures described earlier), or on the overall effect of showcards in a split ballot among face-to-face respondents only.

2) *What are the mechanisms that underpin social desirability bias in telephone interviews?*

While the use of showcards did not appear to effect data comparability in our phase 2 face-to-face and telephone interviews, the differential presence of social desirability bias in each mode was more concerning. In order to minimise the likelihood of social desirability bias in telephone data, it is essential that we understand better the causes and mechanisms underlying the bias and the extent to which they are under the conscious control of the respondent. The commonly-accepted view is that respondents are motivated to execute the response process systematically, but that they edit their true response to the survey question in order to avoid embarrassment (Tourangeau and Smith, 1998; Tourangeau, Rips and Rasinski, 2000). The logical extension of this is that response times to sensitive questions will be longer than those for more neutral questions because the respondent must engage in greater cognitive effort to assess their true response in relation to the social desirability connotations of a question and modify their answer accordingly to portray themselves to the interviewer more favourably. There is some evidence to support this (e.g. Holtgraves, 2004), yet this explanation does not tally with the finding that telephone interviews

[10] Dijkstra and Ongena (2002) used behaviour coding of ESS round 1 pilot interviews and found that questions using showcards were more likely to give rise to confusion among respondents.

are generally conducted at a faster pace than face-to-face interviews and that they carry a greater cognitive burden for respondents, giving them less opportunity to think carefully about their answers. In fact, social desirability concerns may trigger other mechanisms, including self-deception, biased retrieval or even shortcutting. For example, respondents may select the most socially desirable response because it is the easiest, most accessible or salient response available to them (without having to engage in 'deep' processing) – a theory that has been used to explain acquiescent response bias (Knowles and Condon 1999). More research is needed to better understand the processes underlying social desirability bias and if and how they vary depending on the mode of data collection.

Still less is known about the extent to which social desirability bias varies cross-nationally (although, see Johnson and van de Vijver, 2003). An important step towards learning more about this would be to investigate similarities and differences in the social desirability connotations of different questions and topics across countries. Such a process is especially important in relation to attitudinal measures (as opposed to say, measures of sensitive behaviours) and is also critical for establishing the extent to which data are affected by the bias to begin with. A recent study by the US-based survey organisation, Harris Interactive, provides examples of the kinds of methods that would be suitable for tackling this problem (see Frisina et al., 2007)[11].

Traditionally, the solution to the problem of social desirability bias has been to offer the respondent more privacy in the response process (e.g. self-completion methods) to ensure the confidentiality of their responses. However, alternative solutions might present themselves once we understand better the causes of the response effect. For example, it may be possible to minimise the effect by reassuring respondents of the legitimacy of the survey (and about confidentiality issues) – perhaps through an advance letter or some other method (such as a specially-scripted introduction from the interviewer). Research (including a review of existing related studies) will be necessary to identify which methods have the most positive impact on data quality.

To summarise our conclusions relating to the problem of social desirability of bias, the following types of research are recommended:

i) An investigation into the cognitive processes underlying social desirability bias in different modes – particularly in telephone mode. Cognitive interviewing, which allows the researcher to gain insight into the reasons for selecting particular responses would appear to be a method well-suited to this endeavour.

ii) A study of the social desirability connotations of the different topics and questions included in the ESS across participating countries. This could be researched either in the context of the ESS itself, or in a specially-designed survey (e.g. see Frisina et al., 2007; Holbrook et al., 2003).

iii) Comparisons of the effectiveness of different methods for reducing the extent of social desirability bias across different modes (e.g. using split-ballot

[11] Frisina and her colleagues (2007) asked respondents how bad or good a stranger would judge certain actions to be, and how good or bad an impression certain actions would make on a stranger. These were then compared with respondents' reported frequencies of engaging in those actions.

designs to compare different question formats, instructions to respondents, interviewer introductions, and so on).

3) ***Can we identify question types that are more susceptible to mode effects?***

In order to achieve our aim of developing questionnaires that as far as possible are insensitive to the mode of data collection, we need to be able to predict what kinds of errors we expect to see and what types of questions are most likely to be susceptible to errors. This will enable us to implement a range of procedures to minimise the likelihood of certain effects on data quality. Our research so far has gone some way towards achieving this, but more work is necessary so we can develop clear guidelines for researchers designing ESS questionnaires in multiple modes in the future. A brief review of what we know about the relationship between question type and mode effects is provided below to help direct research in the short term towards achieving this aim, by specifying different solutions that could be field tested in experimental studies.

Questions with long lists of response categories – questions of this kind are known to place significant cognitive burden on respondents, leading to different types of errors depending on the mode of data collection. In visual modes (including face-to-face interviews using showcards) long lists of options can lead to primacy effects, whereas in aural modes they can give rise to recency effects. Given what we know, a number of methods present themselves as a way of minimising errors and enhancing the equivalence of measures across modes:

i) Reducing the number of response categories in one or both modes to reduce the difficulty of the task
ii) Changing the layout of visually-presented information to mitigate primacy effects (e.g. randomising or reversing the order of response options is common practice; using multiple lists as opposed to one long list)

Complex questions requiring complex showcards – questions of this kind are known to exhibit mode differences in comparisons between face-to-face and telephone interviewing – either because they are inherently difficult to answer, or because they are impossible to administer over the telephone without the associated showcard (thus requiring substantial changes to the original question format). As we saw in the description of phase 3, potential solutions to this problem include:

i) Asking the complex question as an open question and coding responses into the original face-to-face categories
ii) Developing an equivalent telephone version of the question

Controlled experiments are needed in order to compare distributions of responses collected using different methods and obtain the most comparable alternatives. In some cases, it may be necessary to experiment with eliminating the showcard altogether in the face-to-face mode, in order to enhance equivalence with telephone interviews.

'Sensitive' questions, or questions with social desirability connotations – as noted previously, certain types of questions are more susceptible to social desirability bias than others. Research is needed to identify which topics and questions are most likely to generate socially desirable responses and whether these vary cross-nationally. In addition, experimental designs could be used to field test different methods of reducing the likelihood of respondents giving biased answers. These include:

i) Offering self-completion for certain batteries of items or modules addressing sensitive topics (although this is only feasible in face-to-face interviews, unless telephone interviewing was specifically combined with an SAQ).
ii) Providing instructions to respondents to try to encourage honest answers
iii) Using question formats that have been shown elsewhere to minimise socially desirable reporting

Batteries of scale items – questions presented in a battery using the same response scale can be repetitive for respondents and may lead to certain types of response set, such as digit preference or non-differentiation. As noted earlier, such response strategies are more common in modes that place additional burdens on respondents – such as in telephone mode where there are no visual aids, or in self-completion modes, where the lack of variety may decrease motivation. A further complication is that the visual presentation of the response scale (in modes other than telephone) may generate particular types of effect (including primacy effects, depending on the orientation of the visually-presented response options), whereas the oral presentation of response options may lead to recency effects.

Because the type of problem posed by presenting questions in this way varies with the mode of data collection, different methods of preventing response effects present themselves. In either case, efforts should be made to reduce the burden on respondents and avoid excessive repetition of response formats. Where this is not possible, it may be necessary to experiment with other ways of minimising the repetitive nature of the stimuli, to assess the impact on the extent of response errors. Dillman's (2000) recommendations for designing questionnaires for mixed mode surveys provides some guidance on this and his more recent work on the visual design and layout of self-completion questionnaires is also likely to be informative (e.g. Smyth, Dillman, Christian and Stern, 2006). Similarly, MTMM experiments conducted in different modes would provide guidance on the types of question format associated with the smallest measurement errors (indeed, Saris's work has already gone a long way towards achieving this for face-to-face interviewing – see e.g. Saris and Gallhofer (2007). Future developmental work could apply the same methods to evaluating the quality of different question formats in other data collection modes (in fact, we are currently developing proposals to include such experiments in the phase 4 mode experiment to be funded by the FP7 ESS Preparatory Phase grant).

9. Conclusion

The central coordinating team of the ESS has remained open, in principle, to the possibility of allowing fieldwork in multiple modes in its future rounds, but only if it can be shown that alternatives to face-to-face would obtain comparable data. In this context, the goal of our research has been to find ways of mitigating, or at least of minimising, mode effects on measurement error. This report provided an overview of what we have learned so far and what we still need to do in order to achieve this aim.

Mode effects are caused by a complex interaction between characteristics of the data collection mode, characteristics of the questions being asked and the various 'actors'

(Groves, 1979) involved in the data collection process (notably, the respondent's willingness to expend the effort needed to answer questions accurately and to report their answers truthfully). Our three phases of experimental research have gone some way to improving our understanding of this interaction, but more research is needed to improve our ability to predict the occurrence of mode effects in different contexts, the types of effects likely to occur and to develop the appropriate means to prevent them. The preceding section provided recommendations as to how future research might be best directed in order to achieve this. In particular, we identified a need for research into the effect on data quality of using showcards in face-to-face interviews, research into the cognitive mechanisms underlying social desirability bias, and more experiments to test different solutions to the specific forms of errors associated with particular question types.

There has been much debate about how best to design equivalent questionnaires for use in a mixed mode context. De Leeuw (1992) has argued in favour of optimising the design of a questionnaire for the mode it is to be administered in, in order to minimise measurement error across all modes. By contrast, Dillman (2000) has advocated the harmonisation of questionnaires for mixed mode surveys using 'unimode' design principles to minimise differences in question format across modes. Both approaches have their merits when designing questionnaires for a mixed mode survey from scratch. However, neither is entirely satisfactory in the context of an existing survey – like the ESS – contemplating a switch in mode, where it is undesirable to make changes to the original questionnaire because of the potential interruption this could cause in the continuity of the time series. For this reason, our own efforts to develop mode-resistant questionnaires have been focused on finding questions that are *comparable* with ESS face-to-face measures even though in many cases the simplest solution would be to modify the existing questions. While the recommendations in this report have been developed with these limitations in mind, it will be important to continually evaluate whether this is necessarily the best strategy for the survey in future. In particular, if the ESS is to become a mixed mode survey, we might wish to consider alternative approaches to the development of new modules of questions that deviate from the methods used at present (e.g. using fewer showcards, avoiding questions with long lists of response options and so on). Such approaches might change the look and feel of the face-to-face instrument, but it might also help to enhance comparability across different modes.

Whatever strategy we adopt for developing mixed mode survey instruments, it is clear from this review that we require a multi-pronged approach to the problem of mode effects. Threats to data quality come from a variety of sources, and understanding the mechanisms that underlie them provides the key to how best to mitigate them. The task for the next stage of our preparations for a mixed mode future on the ESS is to exploit as many possibilities for tackling this challenge as we are able.

10. References

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58*(2), 210-240.

Bryson, C., O'Shea, R. and Nicolaas, G. (2003). Report of phase 1 of the ESS-Gallup Mixed Mode Methodology Project. Unpublished manuscript.

de Leeuw, E. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT Publications.

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

de Leeuw, E., & van der Zouwen, J. (1988). Data quality in telephone and face-to-face surveys: A comparative analysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283-299). New York: Wiley.

DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena: Volume 2* (pp. 257-282). New York: Russell Sage.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley Co.

Dijkstra, W., & Ongena, Y. P. (2002). *Question-answer sequences in survey-interviews.* Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing Methods, Charleston, SC.

Duffy, B. (2003). Response order effects - How do people read? *International Journal of Market Research, 45*(4), 457-466.

Epstein, J. F., Ripley Barker, P., & Kroutil, L. A. (2001). Mode effects in self-reported mental health data. *Public Opinion Quarterly, 65*, 529-549.

Eva, G., Roberts, C.E., and Lynn, P. (2008). *Measuring the effect of interview length on telephone survey cooperation*. Paper presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, 15th-18th May 2008

Frisina, L., Thomas, R.K., Krane, D., and Taylor, H. (2007) *Scaling Social Desirability: Establishing Its Influence Across Modes.* Paper presented at the 62nd Annual Conference of the American Association for Public Opinion Research, Anaheim, 17th-20h May 2007.

Groves, R. M. (1979). Actors and Questions in Telephone and Personal Interview Surveys. *Public Opinion Quarterly, 43*(2), 190-205.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews*. New York, NY: Academic Press.

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.

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

Jäckle, A., Roberts, C. E., 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. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2006-41.pdf

Jäckle, A., Roberts, C. E., and Lynn, P. 2008. *Assessing the Effect of Data Collection Mode on Measurement ISER Working Paper,* 2008-08. Colchester: University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2008-08.pdf

Jobe, J. B., Pratt, W. F., Tourangeau, R., Baldwin, A. K., & Rasinski, K. A. (1997). Effects of interview mode on sensitive quesitons in a fertility survey. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons, Inc.

Johnson, T.P. and Van de Vijver, F.J.R. (2003). Social desirability in cross-cultural research. In J.A. Harkness, F.J.R. Van de Vijver and P.Ph. Mohler (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly, 44*, 210-222.

Kaminska, O., Bautista, R. and Serrano, E. (2007). *Best combination of modes*. Paper presented at the 62nd Annual Conference of the American Association for Public Opinion Research, Anaheim, 17th-20h May 2007.

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Applied Psychology, 77*(2), 379-386.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

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., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, 37, 941-964.

Miller, P. V. (1984). Alternative question forms for attitude scale questions in telephone interviews. *Public Opinion Quarterly, 48*, 766-778.

O'Connell, A. A. 2006. *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: Sage.

Peytcheva, E.A., Manchin, R., Tortora, R. and Groves, R.M. (2004). *Comparing face to face, telephone, paper self-administered and web survey measurement*. Paper presented at the 59th Annual Conference of the American Association for Public Opinion Research, Phoenix, Arizona 13th-16th May, 2004.

Roberts, C. E., Jäckle, A., and Lynn, P. (2006). *Causes of Mode Effects: Separating out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys*. Proceedings of the Survey Research Methods Section. American Statistical Association. http://www.amstat.org/Sections/Srms/Proceedings/

Roberts, C. E. (2007). Mixing Modes of Data Collection in Surveys. *ESRC National Centre for Research Methods Review Papers,* NCRM/008. http://www.ncrm.ac.uk/publications/methodsreview/MethodsReviewPaperNCRM-008.pdf

Roberts, C., Eva, G., Allum, N. and Lynn, P. (2008). *Does length matter? The effect of interview duration on data quality in telephone interviews*. Paper presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, 15th-18th May 2008

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.

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

Saris, W.E. and Gallhofer, I. (2007) Can questions travel successfully? In R. Jowell, C. Roberts, R. Fitzgerald and G. Eva (Eds.) *Measuring attitudes cross-nationally: Lessons from the European Social Survey.* London: Sage Publications

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193-212.

Serrano, E., Kaminska, O., McCutcheon, A., and Manchin, R. (2007). *Mixed-Mode Effects: A Multilevel Approach*. Paper presented at the 62nd Annual Conference of the American Association for Public Opinion Research, Anaheim, 17th-20h May 2007.

Smyth, Jolene D., DonA.Dillman, Leah Melani Christian, and Michael J. Stern. 2006. Comparing Check-All and Forced-Choice Question Formats in Web Surveys. *Public Opinion Quarterly* 70:66–77.

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

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275-304.

Tourangeau, R., & 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.

Williams, E. (1977). Experimental comparison of face-to-face and mediated communication: A review. *Psychological Bulletin, 84*, 963-976.

11. Appendix

The phase 1 question wording experiment used a combination of four questions designed to measure net household income:

a) "Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household?" (income sources)
b) "When thinking about your household's net income, do you think in terms of…?" (timeframe)
c) "People's income comes from lots of different sources… if you add up your household's total net income from all sources is it…. Per week/ per month/ per year?" (banded income: telephone only)
d) "Add up your household's total net income from all sources and tick the box next to the appropriate amount. Use the table you know best: weekly, monthly or annual income." (detailed income)

The questions were combined in the following ways in each of the four modes and questionnaire versions:

Face to Face		Telephone		PAPI		Web	
A	B	A	B	A	B	A	B
Questions a & d/ visual stimulus	Question d/ visual stimulus	Questions b, c, &d/ aural stimulus	Questions b, c, & d/ aural stimulus	Questions a, b, & d/ visual stimulus	Questions b & d/ visual stimulus	Questions a, b & c/ visual stimulus	No version B