Measuring the effect of interview length on response propensity and response quality in a telephone survey - Final report of the ESS CATI experiment

ESSi JRA1 Deliverable 5

Caroline Roberts
FORS, University of Lausanne

Gillian Eva
Centre for Comparative Social Surveys, City University, London

Peter Lynn
Institute for Social and Economic Research, University of Essex

Jerry Johnson
Cathie Marsh Centre for Census and Survey Research, University of Manchester

With thanks to the national teams:

Cyprus
Coordinator: Antonis Theocharous
Fieldwork agency: Cyprus College Research Center

Germany
Coordinator: Silke I. Keil and Jan van Deth
Fieldwork agency: infas (Institut für angewandte Sozialwissenschaft GmbH)

Hungary
Coordinators: Agnes Illyes and Robert Manchin
Fieldwork agency: The Gallup Organization Hungary

Poland
Coordinator: Franciszek Sztabinski
Fieldwork agency: Millward Brown SMG/KRC Poland

Switzerland
Coordinator: Dominique Joye
Fieldwork agency: MIS Trend on behalf of SIDOS

Table of contents

1	**Abstract**	2
2	**Background**	2
3	**Introduction**	4
3.1	***Interview length and response propensity***	4
3.2	***Interview length and data quality***	5
4	**Methodology**	8
4.1	***Research Design***	8
4.2	***Sampling***	9
4.3	***Fieldwork***	10
5	**Analysis**	17
5.1	***Response rate***	17
5.2	***Data quality***	18
5.3	***Additional analysis***	22
6	**Results**	22
6.1	***Rates of participation***	23
6.2	***Data quality***	32
7	**Additional findings from the CATI study**	44
8	**Discussion**	46
8.1	***Summary of findings and discussion***	46
8.2	***Problems with the design***	49
8.3	***Recommendations for the ESS***	51
8.4	***Avenues for future research***	53
9	**References**	55
10	**Annexes**	58

1 Abstract

In this paper we report on an experiment carried out in the context of the European Social Survey, designed to examine the effect of inviting respondents to participate in telephone interviews of different lengths on their willingness to participate in the survey. Three treatment groups were interviewed with three versions of the ESS questionnaire, adapted for telephone administration: one the full one-hour questionnaire; one 45 minutes; and one the full ESS questionnaire divided in two roughly equal parts. Due to alterations in the position of one module, we were also able to consider the effect of length on the quality of data collected by comparing the responses in module E in each group. Although not the main aim of this experiment, we also conducted some comparisons between the telephone data and data from the round 3 ESS face-to-face survey in order to identify differences in response rate and data quality by mode of data collection. Overall, we found some differences in response rates and data quality due to both questionnaire length and mode, but the findings were inconsistent. Results differed within each country, suggesting that any decisions on a mixed-mode future should be taken on a country-by-country basis.

2 Background

The Central Coordinating Team (CCT) of the European Social Survey (ESS) has been conducting a programme of research investigating the feasibility of mixing modes of data collection in its future rounds. The aim of this ongoing programme is to provide information that will help to inform decisions regarding:

- whether mixed-mode data collection should be allowed on future rounds of the ESS;
- which modes of data collection might be allowed;
- within which kinds of overall survey design mixed modes could be employed.

The following issues are being assessed:

- coverage and response rates that can likely be achieved with different modes and mode combinations;
- likely differential error between modes (particularly non-response error and measurement error) and its causes.

To date, the research has consisted of a series of experiments focusing on gathering information about mode effects on measurement error. There have been two phases of this research: phase I involved a pilot study conducted in Hungary in 2003, which allowed paired comparisons across all the main modes of data collection (face-to-face, telephone, Internet and paper self-completion); phase II – carried out in Hungary and Portugal in 2005 – was an experiment designed to investigate the likely impact of a switch to telephone interviewing on data quality, in particular looking at the extent and cause of differential measurement error between face-to-face and telephone interviews (see Jäckle et al, 2006; Roberts et al, 2006).

The research undertaken so far has been funded by a modest budget for methodological work, which was built into the contract for the first two rounds of the ESS. It has been carried out in conjunction with Gallup Europe. This collaboration has allowed the CCT to benefit from a larger-scale research project than would have been possible alone, as it has entailed the pooling of financial resources (with both parties contributing 50%). The study reported on here – phase III of the research programme – was funded as part of a Joint Research Activity under the ESS Infrastructure grant (ESSi), which commenced in May 2006.

The study focused on the practical challenges involved in conducting ESS fieldwork by telephone. These include issues relating to questionnaire design, sampling and the selection of target respondents, contact procedures and interview length. The latter formed the primary focus of the research, as it represents the most significant obstacle to administering the ESS questionnaire by telephone.

The average length of the ESS face-to-face interview varies by country, but is estimated to be around one hour. However, telephone interviews are typically designed not to exceed a maximum length due to concerns about break-offs and respondent fatigue. Some survey agencies even try to restrict the duration of telephone interviews to less than 20 minutes. While there is likely to be considerable variation in the tolerance for long interviews by telephone, it was necessary to:

a) establish the extent to which interview length impacts on response propensity and response rates;
b) explore ways of modifying the standard ESS interview to make it more suitable for telephone administration.

Interview length also has implications for data quality, because longer survey interviews (particularly over the telephone) are assumed to place greater cognitive burden on both respondents and interviewers. One possible outcome of increased burden is that respondents will not make sufficient effort to respond to the survey questions systematically, adopting instead what has been referred to as a 'satisficing' strategy in order to reduce the cognitive effort required to answer the questions. Satisficing can take a variety of forms but is generally associated with increased measurement error, and correspondingly, with a reduction in survey quality. A further aim of the research therefore was to:

c) investigate the effect of interview length on response quality.

With respect to (b), in this study the following two options were tested alongside the full hour-long ESS questionnaire:

1. *A modular design* in which different sub-samples of respondents respond to different modules of the questionnaire, thereby reducing the overall length of the interview for all respondents. Given the design of the ESS, it is relatively straightforward to implement such a method, using the rotating modules as the basis for dividing up the questionnaire.
2. *Splitting the interview into two parts* (to be conducted on two separate occasions) – this would have the advantage of reducing the interview time and, thereby, the burden on interviewers and respondents. However, splitting the interview has the disadvantage that respondents may refuse to participate in the second interview, as well as of a possible negative impact on data quality (e.g. from the effects of time between data collection points; from having to re-order questions in the survey or through altering any influence of questionnaire length on response quality and thus disturbing within-country comparisons). In this study, respondents were offered the option of splitting the interview, or continuing to complete the whole questionnaire at one time.

The ESS is currently a uni-mode face-to-face survey and it is likely that any future mixed mode data collection design would continue to include a face-to-face element. Furthermore, four rounds of data collection have already been conducted with face-to-face as the sole mode of data collection. This means that, before adopting an alternative or additional mode, we need to be sure of the equivalence of data between modes. Since the telephone survey was conducted at almost the same time as the round 3 face-to-face fieldwork, using as similar as

possible methodology, it was possible to make some comparisons between the face-to-face and telephone surveys in terms of a) response rates and b) data quality.

3 Introduction

3.1 Interview length and response propensity

Unlike face-to-face interviews, telephone interviews are less suited to the administration of long survey questionnaires. This is partly because the absence of visual cues and greater social distance between interviewer and respondent can impact negatively on data quality (see below). However concerns about carrying out long survey interviews by telephone have mainly been motivated by worries about response rates and possible break-offs mid-interview. The underlying assumption is that the longer the survey interview is expected to last, the less likely it is that a sample member will want to take part – and equally, the longer the interview does last, the less the respondent will want to continue to participate.

Because of this, many survey organizations actually put formal limits on the length of survey interviews to try to encourage participation and discourage break-offs. For precisely the same reasons, and often irrespective of questionnaire length, telephone interviews tend to be conducted at a quicker pace than face-to-face interviews. However, surprisingly few studies have investigated empirically the relationship between questionnaire length and cooperation in telephone surveys.

As Berdie (1973; p.278) has argued, "common sense suggests that the shorter the questionnaire, the more likely a high response rate, and persons studying questionnaire efficiency have tended to accept this belief in spite of little empirical evidence to support it." Even now there is only limited research into the effect of length of questionnaire on rates of refusal for interviewer surveys. Most of the existing evidence relates to mail surveys, for which the relationship between length and refusal rate seems to be much stronger but the issues are quite different. However, there is some evidence that, in face-to-face and telephone surveys, refusal rates are higher in longer interviews. Hansen (2006) found that, when incentives were held constant, an announced 15-minute survey had a 30% greater chance of resulting in a completed survey than a 20-minute survey. Collins and his colleagues (1988; 2001) compared results for a 40-minute survey and a 20-minute survey and found that the longer questionnaire had a 5% higher refusal rate (14% compared to 9%).

However, Collins et al (ibid.) also found that although there were initial differences in response rates for telephone surveys of different lengths, these dropped out when basic follow-up techniques were used. Similarly, a review by Heberlein and Baumgartner (1978) of mail surveys found no difference in the effect of length on the likelihood of participation until topic saliency and number of contacts were controlled for, when the longer interviews did lead to lower response rates. This suggests that if the questionnaire is interesting enough to the respondent and sufficient conversion efforts are made, the effects of length can be overcome.

The assumption appears to be that respondents will be unwilling to take part in long surveys because they pose a bigger burden, but according to Frankel & Sharp (1981) there is not a clear relationship between interview length and respondent's perception of the interview burden. In a study of refusers to a 25-minute survey, only 5% reported that the interview being too long was the reason for refusal (Collins et al, 1988/2001). Morton-Williams and Young (1987) found similar results in a face-to-face survey where only 9% of initial refusers

mentioned length as a problem. Importantly, of those that did mention it, 92% were later converted to respondents. Indeed, Bradburn (1978) suggested that respondents might even view long interviews positively since they could be perceived as being more serious and important. In fact, there is some evidence that at least some of the negative effect of length may not be due to respondent reluctance but instead to interviewer's concerns and expectations of difficulties (Collins et al, 1988/2001). For example, longer interviews create more problems for interviewers in terms of the scheduling and timing of appointments; they cannot be started late in the evening, which may be a good time to reach certain respondents. (Marquis, 1979; Botman and Thornberry, 1992).

A problem with many of the existing studies on interview length and cooperation in surveys is that they have tended to confound various possible causes of refusal. There are many other elements of a survey that will influence likelihood to participate, such as the mode of administration and the topic, and it is hard to disentangle these in an experiment (Groves and Lyberg, 2001). Furthermore, it is difficult to generalise from an experimental study to other surveys with a different design, population, topic, mode and so on.

Here we are also interested in the difference in propensity to respond by mode, in particular comparing telephone response rates with face-to-face response rates achieved on the ESS. There is much evidence that face-to-face surveys achieve higher response rates (de Leeuw, 1992; Holbrook, Green and Krosnick, 2003; Hox and de Leeuw, 1994; Czaja and Blair, 2005), apparently because interviewers find it easier to persuade someone to take part in-person than over the phone. As well as variations in the response rates achieved by mode, different modes also attract different members of the population. Again face-to-face surveys tend to perform the best and achieve fairly equal cooperation across the population (Czaja and Blair, 2005). Telephone surveys, however, have been found to lead to response bias under-representing those with low education, low income, and older respondents (Holbrook, Green and Krosnick, 2003).

Another major advantage of face-to-face surveys is that they offer complete – or at least, very good – coverage (in most countries), whereas telephone surveys will not only exclude those in the population without telephones but also require good quality sample frames with telephone numbers listed, which are not common. Where these lists do exist they *usually* do not include lists of mobile phone numbers so that anybody (or household) with only a mobile phone is excluded. A rise in the proportion of individuals using mobile phones instead of (rather than alongside) fixed-line telephones, especially in Eastern Europe (Blyth, 2007) has led to a rise in problematic undercoverage in RDD surveys. In particular, mobile-only households differ on a number of socio-demographic variables from those with fixed-line telephones, such as age, gender and urbanicity (Roberts, Eva and Widdop, 2008).

Set against these drawbacks of telephone surveys, it is clear that in many countries, response rates on the ESS are falling well below the target 70%, often despite major efforts in response enhancement. Furthermore, the samples achieved are not always representative of the population (Roberts, Eva and Widdop, 2008; Billiet and Meuleman, 2004; Vehovar and Zupanič, 2007). The need to explore alternative modes of data collection, including mixed mode designs, has arisen from these challenges.

3.2 Interview length and data quality

As mentioned, length of the interview can have implications for the quality of survey data. Survey errors are wide-ranging and can be difficult to predict and measure. Nevertheless, there is now an extensive literature documenting the different types of response errors that can

affect the overall quality of survey data, showing consistent patterns about when and where such errors are likely to occur. There are a number of different approaches to examining measurement error in surveys, discussed below.

Groves (1979) has argued that measurement error in surveys can be attributed either to the 'actors' involved in the survey process (notably, the interviewer and respondent in interviewer-administered surveys) or 'questions' (or how the questionnaire is administered to survey respondents). Tourangeau, Rips and Rasinski (2000) categorise the various cognitive processes involved in answering survey questions into four main components of processing (each consisting of several sub-components): (1) comprehending the survey question, (2) searching for and retrieving from memory the information requested, (3) formulating a judgment based on the retrieved information, and (4) mapping that judgement on to the available response options in order to select and report an answer. Problems can arise during any of these various processes, leading to errors in the data.

The 'questionnaire satisficing' approach developed by Krosnick (1991) provides an explanation as to why respondents' answers may contain errors. According to this approach, executing each of these stages of processing carefully represents the 'optimal' approach to survey responding and many diligent, conscientious respondents may indeed participate in surveys in this way. However, it is likely that for many respondents, the cognitive effort required to complete each of these processes systematically will outweigh the motivation needed to do so. As a result, respondents may – consciously or unconsciously – take shortcuts to reduce the amount of cognitive work involved in the survey task. These shortcuts may take the form of going through each of the necessary processes, but only doing so superficially (referred to as 'weak satisficing'), or it may take the form of skipping processes out altogether (referred to as 'strong satisficing'). Different types of errors may be observed, depending on the nature of the shortcutting. For example, weak satisficing includes response effects such as 'acquiescence', a bias towards agreeing with assertions in the question, and 'response order effects' which arise when respondents select the response category that is most accessible in memory – either at the start of a list, where the options are presented visually or at the end of a list where the options are presented orally (Krosnick and Alwin, 1987), while strong satisficing includes effects such as repeatedly selecting the 'Don't Know' option, and 'non-differentiation', in which items to be rated on the same response scale are rated on the same scale point (see Krosnick, 1991; Krosnick, Narayan and Smith, 1996; Krosnick, 1999). Other response strategies have also been investigated as possible indicators of satisficing, including selecting the middle response category and 'extremeness', a preference for selecting answers from the end points of a scale (Holbrook, Cho and Johnson, 2006).

The likelihood of respondents adopting a sub-optimal or satisficing response strategy depends on the respondent's ability to engage in the necessary processing, their motivation to do so, and the difficulty of the survey task itself. A large number of studies provide evidence consistent with this model (e.g. see Krosnick, Judd and Wittenbrink (2005) for a review). Each of these factors may be further influenced by other variables in the survey setting. For example, the respondent's ability to expend the required effort may be affected not only by individual factors, but also situational ones, such as the presence of distraction. Motivation to respond 'optimally' may be influenced by the nature of the survey topic, whereas task difficulty (i.e. the cognitive burden of completing the questionnaire) will depend not only on topic, but also on factors such as the types of questions asked, the complexity of question wording, and so on.

Response effects found in survey data are assumed to have resulted from questionnaire satisficing where they occur under 'conditions that foster satisficing' (Krosnick, 1991) – notably, where respondent ability and motivation are low and task difficulty is high. Two variables that have been found to influence these conditions are of particular interest here: the mode of data collection and the length of the survey questionnaire. In relation to mode, face-to-face interviewers are better able to keep respondents engaged in the survey task because they can react quickly if the respondent's motivation appears to be flagging. Equally, they are better able to pace the interview to suit individual respondents' needs (e.g. de Leeuw, 1992). By contrast, telephone interviews typically place a greater cognitive burden on respondents because they are often conducted at a faster pace, and because of the absence of visual cues that can be used to aid respondent concentration and recall. This is particularly important on a survey such as the ESS which has questions with long, complicated response options and relies heavily on showcards.

Different satisficing behaviour may be more likely in one mode than another. For example, in self-completion modes, questions using the same response scale presented as batteries might encourage non-differentiation. In modes where the response options are presented orally, respondents might be more likely to select the last response option they are offered due to the high cognitive burden of remembering the whole list. Alternatively, in modes where the response options are presented visually, respondents might be more likely to select the first response option, to avoid the effort of reading the whole list (Krosnick and Alwin, 1987). Overall, however, questionnaire satisficing is predicted to be more likely in telephone interviews than in face-to-face interviews and there is some evidence to support this (see Holbrook, Green and Krosnick, 2003 for an overview). Holbrook and her colleagues found more evidence of satisficing among telephone respondents than among respondents interviewed face-to-face. However, Jäckle, Roberts and Lynn's (2006) comparison of the two modes found little evidence of differences in satisficing effects across modes.

The contrast in these findings seems likely to have resulted from differences in the length of the survey questionnaires in each study. Holbrook et al.'s research compared responses to the National Election Study (NES), where the interviews lasted around 1 hour in total. The study reported by Jäckle et al., used questionnaires containing just a subset of items from the European Social Survey and the interviews only lasted around 15 minutes, so they were presumably less burdensome for the respondents than the NES interviews. Longer questionnaires are predicted to be more likely to encourage satisficing because the respondent's motivation typically wanes as he or she progresses through the items. As the respondent tires, ability to concentrate is also likely to decrease and correspondingly, cognitive burden increases, making shortcutting more likely. Consistent with this, researchers have found evidence of more satisficing on items placed towards the end of the questionnaire (see Roberts, Eva, Allum and Lynn, 2008, for a review), however few studies have explicitly attempted to compare data quality across interviews of different lengths to test the hypothesis that longer questionnaires are more susceptible to response effects than short questionnaires (although see Herzog and Bachman, 1981).

It is likely that there are considerable cross-national variations in tolerance for long survey interviews by telephone, depending on variations in survey practices across countries and individual survey climates, which may influence the effect of length on both response rates and data quality. In addition, the survey climate in different countries can mean that the effect of mode of data collection differs across countries. These assumptions are certainly underpinned by anecdotal evidence (see Roberts, Eva and Widdop, 2008), although again, there do not appear to be many documented empirical comparisons of this (but see Groves and Lyberg, 2001; p.206). Such cultural differences can have implications for the design of

comparative surveys: if a single mode data collection strategy is to be adopted (which offers considerable advantages in terms of enhancing the equivalence of the data across countries), then the questionnaire needs to be designed to an optimal length that will ensure it can be successfully administered in all participating countries, without causing wide variation in response rates (which have their own implications for comparability) and data quality. If a mixed mode strategy is to be used, then it is important that a questionnaire that works well in one mode will also work well in another mode, without seriously affecting data comparability within each country.

4 Methodology

In this paper we report an experiment that was designed to explore the feasibility of allowing a switch from face-to-face to telephone interviewing. The experiment was designed to test whether it would be possible to run the whole ESS questionnaire by telephone (which takes around 1 hour to administer face to face) as well as to test possible alternatives if the full hour-long questionnaire proved impossible. This design allowed us to examine the effect of varying the length (and structure) of the interviews on rates of participation (by informing target respondents during the survey introduction of the anticipated duration of the interview). In addition, since varying the interview length required some structural changes to be made to the questionnaire, the design also allowed us to examine the effect of interview length on data quality, primarily measured by propensity to satisfice, using a module for which the placement differed in each version. Finally, because the experiment was run at (roughly) the same time as the round 3 face-to-face fieldwork, we were able to make some comparisons between the two modes, with regard to data quality and rates of participation.

The survey experiment was carried out in five countries, all of which participated in round 3 of the ESS (2006/2007): Cyprus, Germany, Hungary, Poland and Switzerland. The selection of countries was based mainly on pragmatic considerations and the available budget, but within those constraints, we chose countries with divergent traditions of survey practice, facing different challenges in their data collection efforts on the ESS (mainly related to the costs of face-to-face interviewing and the response rates obtained in that mode). Sample members (selected using strict probability sampling methods in each country) were randomly assigned to one of three treatment groups, which varied according to the length and design of the questionnaire. The estimated interview lengths were as follows: group A) 60 minutes; group B) 45 minutes; and group C) 2x30 minutes. Interviewers were instructed to tell the selected target respondent during the survey introduction how long the interview was likely to be (30 minutes in the case of group C).

4.1 Research design

The ESS face-to-face questionnaire consists of 4 core modules of questions repeated at each survey round and two 'rotating modules' covering new substantive topics that are unique to each round (though they may in future be repeated). The core questionnaire covers a range of topics including media consumption, social and political trust, political interest and participation, religious and ethnic identity and socio-demographic characteristics of the respondent and his or her partner and parents. In round 3, one of the rotating modules measured attitudes towards the timing of life events and transitions from youth into adulthood and old age, and the other was designed to measure personal and social wellbeing. The whole questionnaire takes around one hour to be administered face-to-face (with some variations between countries).

The questionnaire used for respondents in group A of the telephone experiment was essentially identical to the face-to-face version, with some adaptations that were needed to

make it suitable for a telephone interview (see section 4.3.5). The questionnaire used in group B included just one of the round 3 rotating modules, personal and social wellbeing (reducing the overall length by 55 items), and was estimated to last around 45 minutes. The rationale behind this design was that one way to shorten the questionnaire would be to ask each rotating module of just half the sample (for this reason the sample size for group B was half that of groups A and C). In group C, sample members were asked to take part in an interview lasting around 30 minutes, at the end of which they were asked if they would be willing to complete another interview of the same length either straight away or another time. See section 4.3.4 for details on how version C was split. The design is summarised in table 1, along with the proportion of the issued sample allocated to each treatment group. Mean interview lengths in each group for all countries combined confirmed the initial estimates. (Analysis of length by country can be found in section 6.2.1).

Table 1 – Research design

Group	Treatment	Estimated interview length	Issued sample %
Group A	Full ESS interview	≈ 60 mins	*40*
Group B	Full ESS interview –55-item module	≈ 45 mins	*20*
Group C	Full ESS interview in two parts	≈ 2*30 mins	*40*

Due to the removal of one module in group B and the change in order required in group C, module E (the rotating module on personal and social wellbeing) appeared at a different point in each questionnaire version. Table 2 shows the number of items preceding module E in each version. In version A it was preceded by 141 items, in version B by 86 items, for version C participants who completed the two parts of the interview in one go it was preceded by 166 items, and for version C participants whose part 2 interview took place on a separate occasion to their part 1 interview it was preceded by just 36 items. To examine the effect of varying the length of the questionnaire on data quality, we decided to focus on responses to this particular module. (Question wording and response options are shown in Annex 1).

Table 2 – Position of module E

Version	Number of items before E
Version A	141
Version B	86
Version C in 1 part	166
Version C in 2 parts	36

4.2 Sampling

The ESS is intended to cover individuals aged 15 and over (no upper age limit) resident within private households in each country, regardless of their nationality, citizenship or language. In this study, resource constraints restricted us to relatively small sample sizes, but participating fieldwork agencies were instructed to use the best possible probability sample design available (in all cases this was developed in consultation with one of the authors, who is a member of the ESS panel of sampling experts). As with the ESS, sample designs were allowed to vary cross-nationally, depending on the availability of sampling frames in each of the different countries. In all cases the samples selected were of households, so at the first contact interviewers were required to use a random selection procedure to identify a target respondent and no substitutions were allowed (in line with the mainstage face-to-face specifications).

In the end, three out of the five countries used list-assisted methods of RDD sampling. Cyprus used an electronic catalogue of numbers provided by the Cyprus Telecommunications Authority, which includes listed and unlisted fixed line telephone numbers and details of the district (urban/rural), but no names or addresses. It is understood to provide full coverage but includes shops and small businesses as well as households (around 8-10% of numbers were estimated to be ineligible). Switzerland used the Swiss telephone directory, which is understood to provide a level of coverage of around 97% of resident households. Note that in Switzerland the sample was restricted to the French-speaking population. In the remaining countries, the samples were intended to represent the ESS population, though in practice, were representative only of those households with fixed-line telephones. The proportion of cell-phone only households varies widely in Europe, but is estimated to be around 6% in Germany, 33% in Hungary, 22% in Poland and 1% in Switzerland (see Roberts, Eva and Widdop, 2008) [1]. Given the aim of the study was to examine relative differences between the three experimental groups (to which participants were randomly assigned), rather than to make inferences to the population as a whole, the resulting under-coverage was not deemed to be overly problematic (though of course it has considerable implications for the suitability of telephone interviewing for the ESS more generally). Where comparisons were made between mode, in the face-to-face data only those with fixed-lines in their accommodation were included and for the Swiss data only the French-speaking regions, to bring the sample in-line with the telephone sample.

Each country was instructed to start with a probability sample of around 1000 *eligible* cases and to use a procedure for randomly assigning sample members to one of the three experimental groups, proportionate to the issued sample sizes specified in the sample design.

4.3 Fieldwork

4.3.1 Fieldwork specifications

While this study is focused on different methods of carrying out the ESS by telephone, one of the main purposes of the overall research was to make comparisons with the standard face-to-face method used in each participating country, in particular with regard to response rates. For this reason, the fieldwork procedures used to implement the experiment were matched as closely as possible to those being used for round 3 of the ESS in each country. For example, the timing of fieldwork was as close as possible to the face-to-face data collection period. Fieldwork was carried out either during or just after the ESS round 3 data collection period between November 2006 and February 2007. However, certain changes to the standard procedures were necessary due to the fundamental differences between the face-to-face and telephone modes of data collection. Survey agencies were requested to optimise fieldwork procedures to the standards of best practice for telephone interviewing. As with the main ESS, the experiment aimed to meet the highest methodological standards. In order for the information gathered to be truly comparable, it was essential that equivalent methods were used in all participating countries and so detailed project specifications were developed by the project team and provided to national co-ordinators in the participating countries, who liaised with survey agencies on our behalf.

The precise fieldwork procedures used in each country varied depending on the available budget (e.g. only Switzerland used an advance letter and incentives – see section 4.3.2). After completing the selection procedure and making contact with the target respondent,

[1] Data from ESS round 3 (2006), edition 3.2 Note there are differences in how the estimates are derived in different surveys and that these figures may change rapidly (see Roberts, Eva and Widdop, 2008).

interviewers were instructed to introduce the survey (as the ESS) and to include in their introduction an estimate of the likely length of the interview[2], depending on which group the respondent had been randomly assigned to. Note that for group C respondents the survey was introduced initially as a 30-minute interview (for group A as 60 minutes and group B as 45 minutes). At the end of the interview, respondents in group C were then asked whether they would be willing to participate in a follow-up interview, also of 30 minutes. In all cases, interviewers were encouraged to try to arrange appointments for respondents to participate in the survey at their convenience.

Fieldwork agencies and interviewers were also given strict instructions to record all call attempts made to sample units and target respondents. Call outcome codes were provided that conformed to recommendations for recording final dispositions in telephone surveys by Lynn and his colleagues (2001)[3] and which were broadly equivalent to the data collected on the contact forms for the face-to-face ESS, making it possible to make comparisons both across the three experimental groups and across the two modes in the calculation of response rates. Two separate lists of outcome codes were used depending on whether the sample was RDD or from a list of telephone numbers/ households (both are shown in annex 2).

For each call, the following data items were collected:

1. Serial Number of the Sample Phone Number
2. Date of call (DD/MM)
3. Time of call (HH/MM)
4. Call outcome code (see annex 2)
5. Interviewer identification number
6. Number of persons aged 15+ in household (if respondent selection made on this call)
7. Answers to 4 additional questions (in the case of a refusal)

As is the case on the standard face-to-face ESS, field agencies were instructed to make a minimum number of call attempts (10) to each number sampled before closing the case as a 'non-contact'. They were also instructed that 2 of these calls were to be made in the evening and 2 at the weekend and that they should be spread across the fieldwork period (which itself was a minimum one-month period and a maximum of two months) to maximise the chances of making contact with the sampled case.

As with the main ESS, a minimum target response rate of 70% was specified. The maximum proportion of non-contacts specified was 3% of all sampled units[4] for named samples (of individuals and households) and 10% for RDD samples.

Guidance was given on what should be included in the interviewer's introduction to ensure some consistency in how the study was introduced in each country. In particular interviewers were to mention an estimate of the length of the interview, which would vary depending on which experimental group the target respondent was allocated to.

[2] In Switzerland an estimate of interview length was included in the advance letter (varied by treatment group).
[3] Lynn and his colleagues from the UK's Office for National Statistics and the National Centre for Social Research were charged with devising a recommended list of standard outcome categories and definitions for response rates in social surveys carried out in the UK. Though they draw on the AAPOR standard definitions (AAPOR 2006), they identify several reasons why they are not directly applicable to the UK context (and other European surveys), notably due to differences in the nature of sampling methods and sampling frames used in social surveys in the US compared with in the UK and elsewhere in Europe.
[4] The acceptable ratio of non-contacts to sampled units will depend on the sampling method to be used.

Finally, as in the mainstage ESS: a limit was set on the number of interviews each interviewer should conduct; all interviewers were to be briefed; 10% of all calls were to be monitored; and back-checks were to be conducted on 5% of all refusals.

In summary, the fieldwork for this study involved the following procedures:

1. Dispatching an advance letter to each sampled address or individual, where this is standard procedure for the ESS in each country and where budget permitted.
2. Interviewers making contact with all issued phone numbers for households or individuals, and recording specified outcome information for every call made (minimum of 10 calls to each number before accepting a "non-contact").
3. For samples of phone numbers or households, selecting one adult (aged 15 or over) for interview.
4. Introducing the ESS, explaining the procedures for interviewing (including the estimated length of the interview depending on which group the respondent had been allocated to) and, if necessary, arranging an appointment to interview the target respondent.
5. Conducting the interview with the target respondent.
6. Carefully recording the start and end time of each interview (including for both parts of C).

4.3.2 Fieldwork documents

The following fieldwork documents were provided to the survey agencies:

- Project instructions
- 3 versions of the questionnaire (1 for each mode)
- Guide to translation
- Guide to call records
- Instructions on data preparation
- Technical Summary Form

4.3.3 Participating Countries

As mentioned above, the survey experiment was carried out in Cyprus, Germany, Hungary, Poland and Switzerland, all of whom also took part in round 3 of the ESS. While we aimed to choose countries that represented the main areas within Europe, where we might expect survey practices and challenges to be similar, we were somewhat restricted by logistical considerations and available budget. Where possible the same survey agency was used for the telephone experiment and the Round 3 face-to-face fieldwork. Table 3 below shows the survey agencies used in the experiment. Table 4 shows the fieldwork costs.

The field agency in Cyprus experienced several setbacks during the data collection process, including problems with their CATI system, which ultimately led to a decision to terminate the fieldwork period prematurely. There were also serious problems with the face-to-face fieldwork as a large number of selected units were dropped near the end of fieldwork. Consequently the results form Cyprus will not be discussed in this report.

Table 3 – Survey agency

Country	Survey Agency	Same survey agency as R3 F2F fieldwork?
Cyprus	Cyprus College Research Center	Yes
Germany	infas (Institut für angewandte Sozialwissenschaft GmbH)	Yes

Hungary	The Gallup Organization Hungary	Yes
Poland	Millward Brown SMG/KRC	No
Switzerland	MIS Trend on behalf of SIDOS (Swiss Information and Data Archive Service for the Social Sciences)	Yes

Table 4 - Fieldwork costs (*REDACTED)*

Comparing the total survey costs across countries (as shown in table 4), is of limited use, other than to comment that there are much larger differences across countries in cost for the face-to-face survey (range=422,983) than for the telephone survey (range=18,630), presumably because of difference in labour costs. As we found in the mapping report, in those countries where face-to-face is particularly expensive, it also tends to be a lot more expensive than the next most expensive mode. In countries where face-to-face fieldwork is not as expensive, the difference in cost by mode is smaller (Roberts, Eva and Widdop, 2008). As a result, we would expect those countries where data collection is the most expensive to be more interested in switching to an alternative, cheaper, mode.

Costs per interview (that is, the total survey costs divided by the number of valid, complete interviews) tell quite a different story than the total survey costs. In Poland the costs per interview were the same in both modes, and in Switzerland and Germany the cost of a telephone interview was almost half that of a face-to-face interview.[5]

[5] Future analysis should consider fieldwork costs in relation to the amount effort (contact attempts) for each sample member.

Table 5 - Date of data collection periods

Country	Face-to-face		Telephone	
	Dates	Length	Dates	Length
Cyprus	02.10.06 – 10.12.06	10 weeks	29.11.06 – 15.12.06	2.5 weeks
Germany	01.09.06 – 15.01.07	19.5 weeks	05.01.07 – 04.02.07	4.5 weeks
Hungary	21.11.06 – 28.01.07	9.5 weeks	16.01.07 – 06.03.07	7 weeks
Poland	02.10.06 – 13.12.06	10.5 weeks	14.12.06 – 10.02.07	8.5 weeks
Switzerland	24.08.06 – 02.04.07	31.5 weeks	21.11.06 – 28.02.07	14 weeks

Table 6 - Use of advance letter

Country	Face-to-face	Telephone
Cyprus	Yes	No
Germany	Yes	No
Hungary	Yes	No
Poland	Yes	No
Switzerland	Yes	Yes

Table 7 - Use of incentives

Country	Face-to-face	Telephone
Cyprus	Yes (conditional non-monetary)	No
Germany	Yes (conditional monetary)	No
Hungary	No	No
Poland	Yes (unconditional non-monetary)	No
Switzerland	Yes (conditional monetary)	Yes (conditional monetary & non-monetary)

Tables 5 to 7 above show that fieldwork procedures were not always as comparable between face-to-face and telephone as we would have liked, primarily due to cost restraints. However, there are also logistical differences in what is possible in telephone surveys and face-to-face surveys. In all countries the fieldwork periods either overlapped or the telephone fieldwork was conducted directly after the face-to-face fieldwork. In Germany, Hungary and Poland the telephone fieldwork lasted between 1 and 2 months as specified, while in Switzerland it lasted over 3 months. In all cases it was shorter than the face-to-face fieldwork period.

Although all countries used an advance letter in the face-to-face survey, only one country (Switzerland) did so in the telephone survey. This may be because addresses were not available on the sampling frame used. This could be a problem if CATI were introduced on the ESS because we know that advance letters can have positive effects both on the response rates and for interviewers contacting respondents. All countries except Hungary use incentives in the face-to-face survey, but only one (again Switzerland) did so for the telephone survey. As with the advance letter, this might be partly because addresses were not available and so incentives could not be sent in advance. It is possible that the length of data collection

period, and use of advance letter and incentives would have made a difference to the response rates achieved, but we do not address this in the analysis presented here.

In the analysis we present here, we did not conduct an extensive analysis of the call record data from each country. However, further exploration of these data would allow us to evaluate whether participating countries adhered to the protocol, and whether the response rates might have been improved by adjusting the specification, e.g. by insisting on a greater number of contact attempts or extending the fieldwork period'

4.3.4 Questionnaires

The questionnaire used for this study was adapted from the ESS Round 3 face-to-face questionnaire, which included the following sections:

Table 8 - Questionnaire sections

Section		Q#	Topics
A	Core	A1 –A10	Media; social trust
B	Core	B1 – B40	Politics, including: political interest, efficacy, trust, electoral and other forms of participation, party allegiance, socio-political orientations
C	Core	C1 – C36	Subjective well-being, social exclusion; religion; perceived discrimination; national and ethnic identity
D	Rotating module	D1-D55	Timing of life; the life course; timing of key life events, attitudes to ideal age, youngest age and oldest age of life events, planning for retirement
E	Rotating module	E1-E55	Personal and social well-being, helping others, feelings in the last week, life satisfaction, satisfaction with work.
F	Core	F1 – F73	Socio-demographic profile, including: household composition, sex, age, type of area, education & occupation of respondent, partner, parents, union membership, income, marital status
G	Supplementary		Human values scale
H	Supplementary		Test questions
I	Interviewer questionnaire		Interviewer self-completion questions

The supplementary questionnaire (sections G and H) were not used in the experiment. However, interviewers were asked to complete the questions in section I. In addition, five new questions were added to the end of the questionnaire, specifically for the purposes of this study (see Annex 3). Table 9 shows which sections were included in each of the 3 versions of the questionnaire. Version A of the questionnaire was identical to the face-to-face questionnaire for Round 3 (with adaptations for telephone administration described below).

Version B was identical to version A but with section D removed. Version C was more complicated because it was designed for a two-part interview. There were three key considerations when splitting up the questionnaire for version C: firstly to include the key socio-demographic questions in the first part to ensure that the data contained sufficient background variables to be useful to analysts even if the respondent refused to participate in the second part; secondly to ensure that both parts were roughly the same length; and finally to try to ensure that respondents would find the first part interesting enough to want to continue to the second part. As a result of these changes, the order of the sections was modified and section F was divided into two parts (called section X and section Y).

Table 9 - Questionnaire design for the telephone experiment

Group	Questionnaire	Questionnaire design
Group A	Version A	Full ESS questionnaire modified for telephone interviewing + 5 new questions & section I (interviewer questions)
Group B	Version B	Sections A, B, C (core) + section E (rotating module on personal and social well-being) + section F (core) + 5 new questions & section I
Group C	Version C	Part A: Sections A, B (core) + section D (rotating module on timing of life-course) + 25 questions from section F (core) labelled section X Part B: Section C (core) + section E (rotating module) + remaining questions from section F (core) labelled section Y + 5 new questions & section I

4.3.5 Adapting the face-to-face questionnaire for telephone administration

The face-to-face questionnaire had to be amended to make it suitable for telephone administration, the main difference being that no showcards were used in the telephone interviews. Most questions needed no, or very minor, changes but a few required more substantial adaption. Participating survey agencies were requested to use the Round 3 questionnaire that had been translated and prepared for face-to-face fieldwork in their country and were provided with precise details of the changes to be made. The following provides a brief summary of how the telephone versions differed from the standard face-to-face version:

1. All references to showcards were deleted from the questionnaire. For most questions, the absence of showcards meant that the interviewer had to read out the list of response categories to respondents.
2. New interviewer instructions were added to instruct interviewers where they should read out response options, or code open responses.
3. Questions with long lists of response categories or complex showcards were modified (e.g. frequency of behaviours like watching TV). These questions were problematic because the lists of response options are often long and complicated to read out. For some of these items the telephone version became an open-ended question. In other cases, the number of response options was reduced to ease administration by telephone. In other cases the questions were broken down into two or more separate questions.
4. Demographic questions were modified. A number of important socio-demographic measures on the ESS rely on detailed showcards to help the respondent classify

themselves and their relatives (in terms of their educational qualifications, their socio-economic status and so on). These questions tended to require the most modification so that they could be administered without the need for showcards.

Details of the questions in each of the above categories can be found in Annex 4.

5. Analysis

5.1 Response rates

Although it is widely accepted that telephone interviews are not well suited to long questionnaires there is limited empirical evidence to support this. In this study, due to the detailed call records kept by fieldwork agencies, we were able to make a detailed assessment of the effect of the experimental treatment (interview length) on survey cooperation. Our hypothesis was that response rates would be highest for the shorter questionnaires.

As well as looking at how the announced length of interview affected overall levels of cooperation and rates of refusal to participate, we were also interested in whether or not the actual length of the questionnaire influenced the likelihood of respondents to break off mid-interview. Finally, we considered some of the implications of the relationship between interview length and nonresponse for nonresponse bias, by looking at the socio-demographic composition of the samples across the three groups.

As well as the effect of length of interview on response rates and response bias, we were also interested in the effect of *mode* and so we compared the response rates achieved for version A of the telephone experiment with those achieved in the Round 3 face-to-face survey. Our hypothesis is that the face-to-face surveys will achieve higher response rates than the telephone surveys. And to examine the effect of differential response rates by mode on nonresponse bias we also compared the socio-demographic composition of the samples across modes. Only version A was included in these analyses because it was the same length (in terms of number of items) as the face-to-face questionnaire. In the face-to-face data all cases without fixed-line telephones were removed and for the Swiss data, we only included French-speaking respondents from the regions included in the CATI study.

Response rates are typically derived from the final disposition of each case in the survey sample (in this study of each telephone number sampled). The final disposition code can either be calculated as the outcome of the last call attempt to the sampled number/address/individual or by using priority coding to determine which of the calls provides the most accurate description of the cases' disposition. For the telephone data, in this report we present the outcome of the most recent call as the final disposition. Although priority coding may be more informative, the differences between the two different approaches have been found to be quite small in practice (McCarty, 2003; Philippens et al, 2003).[6]

In order to compare the three treatment groups in terms of outcomes, we report on the following outcome measures, based on recommendations by Lynn and his colleagues (2001; pp.43-48) and AAPOR's Standard Definitions (2008; pp 34-38):

1) **Full response rate (AAPOR RR1)** – defined as the number of cases *completing* an interview (i.e. the overall cooperation rate minus any partial interviews/ break-offs), divided by the number of eligible cases in the sample. Note that for cases assigned to

[6] It is the aim of our future analysis of these data to establish the extent of these differences by assigning final outcome codes according to priority coding.

group C, a complete interview is defined as a case completing both parts of the interview – in other words, it excludes those who refused to complete part 2 of the interview. This is the equivalent of the ESS response rate.

2) **Overall response rate (AAPOR RR2)** – defined as the number of cases resulting in an interview (whether complete or partial), divided by the total number of eligible cases in the sample.
3) **Full cooperation rate (AAPOR COOP1)** – defined as the number of complete interviews divided by the number of eligible sample units *contacted*.
4) **Overall cooperation rate (AAPOR COOP2)** – defined as the number of complete and partial interviews divided by the number of eligible sample units contacted.
5) **Household contact rate (AAPOR CON1) –** defined as the number of contacts (interviews, refusals and other contacts not resulting in an interview) with the sampled household, divided by the number of eligible cases in the sample.
6) **Respondent contact rate** – defined as the number of contacts with the selected *target respondent* (interviews, refusals and other contacts not resulting in an interview), divided by the number of eligible cases in the sample.
7) **Refusal rate (AAPOR REF1)** – defined as the total number of cases resulting in a refusal (of any kind – i.e. including refusals by proxy), divided by the number of eligible cases in the sample.

Lynn and his colleagues (2001) have argued in favour of weighting response rate estimates in order that they may properly "reflect the structure of the survey population" (p.45). Because our concern here was with the relative differences between the treatment groups, we present unweighted response, contact and refusal rates for the telephone experiment and face-to-face survey. For the telephone experiment, cooperation rates (also unweighted) are also presented in order to measure the proportion of partial interviews and break-offs. Although we would have been interested to know whether likelihood to break off mid-way through the survey differs by mode, these data are not available for the face-to-face survey because insufficient information is available about partial interviews to calculate the cooperation rates. Note also that we have not dealt here with the issue of unknown eligibility. For the purposes of this paper, we have assumed that all cases in the telephone experiment where eligibility was not established (i.e. all cases for which every call attempt resulted in a non-contact in countries using RDD sampling) were eligible to participate in the survey. Only sample units for which ineligibility was confirmed by the interviewer were discounted from the base for the purpose of calculating outcome rates. For the face-to-face survey, eligibility tends to be easier to calculate because the interviewer visits the address or household and can assess its eligibility.

5.2 Data quality

Data quality was measured in a number of ways in this study: similarity of response distributions across the three treatment groups; item non-response and the extent of respondent satisficing. In all cases, our primary hypothesis was that data quality would vary as a function of questionnaire length. In particular, we expected to see more evidence of satisficing in longer interviews.

This hypothesis (see Krosnick, 1991; 1999) is based on the idea that respondents are more likely to shortcut the response process when motivation is low and task difficulty is high (and particularly where ability to execute the cognitive processes systematically is limited). Motivation is likely to wane over the course of a long questionnaire, while response burden is likely to increase, so we would expect satisficing to be more likely to occur towards the end of the interview, if it is to occur at all (Jabine et al., 1984; p.19; Krosnick 1991; p.224).

For the purposes of the present study, we focused on four response effects that have been repeatedly shown to be consistent with the theory of survey satisficing, in that they are more common and stronger where there is greater task difficulty, lower respondent motivation and among respondents with less education. The evidence relating to the specific effects of questionnaire length are somewhat mixed, however, though in general they lend support to our hypothesis that response effects will be more likely to affect data from longer survey questionnaires.

i) *Non-differentiation* – The tendency to select the first reasonable response for the first item in a set and then use the same scale-point to rate all (or most) of the remaining items in the set (Krosnick, 1991; p.219). Non-differentiation has been observed more frequently on sets of items placed later in the questionnaire (e.g. Herzog and Bachman, 1981; Kraut, Wolfson and Rothenberg, 1975) and has been found to be more common among respondents with less education (e.g. Krosnick and Alwin, 1989).

ii) *Acquiescence* – The tendency to agree with assertions in dichotomous questions, irrespective of their content. Numerous studies provide evidence that the bias results from satisficing, but the findings relating to our specific hypothesis about questionnaire length have been less compelling (e.g. Clancy and Wachsler, 1971).

iii) *Preference for middle alternatives* – The tendency to select the neutral or noncommittal response option in a rating scale (Krosnick, Judd and Wittenbrink, 2005; p. 37). Some research supports our hypothesis of greater use of middle alternatives towards the end of questionnaires (Herzog and Bachman, 1981), but others have found mixed results (Narayan and Krosnick, 1996).

iv) *Response order effects on rating scales* – The tendency to select the first or last response category in long lists of alternatives. Primacy and recency effects are more common among respondents with less education (see Krosnick and Alwin, 1987) but the direction of the effect observed is not always easy to predict – particularly where rating scale items are concerned (as opposed to questions with long lists of response categories).

As discussed, the effect of length on data quality will be tested on questions in module E, which appeared in a different position in each version of the questionnaire. Note that it is the number of items preceding module E that is important for this analysis, as opposed to the full length of the questionnaire. Module E appears latest in version A, then B, then C (when conducted in two parts) (see table 2). This module is also well-suited for our purposes because response effects associated with satisficing are more likely to manifest themselves across response scales, which can often feel repetitive for respondents. Equally there were weaknesses of the choice of module, including the topic, which may have, in fact, motivated respondents, and the possibility of introducing question order effects by changing the position of the module in the questionnaire as a whole. We return to this limitation of the design of our study in the Discussion.

To measure data quality across the three treatment groups, we looked at the similarity of marginal distributions on each of the items in the module, using Chi-Square tests to identify statistically significant differences between groups. To assess item non-response between the groups, we compared the mean proportion of items in the module for which the respondent had given either a refusal, a 'Don't Know' response, or for which there was simply no recorded data (a missing value coded 'No answer')[7]. The main focus of our analysis,

[7] Note that 'Don't Know' is not explicitly offered as a valid response in the ESS questionnaire, but interviewers are instructed to record all refusals and 'Don't Knows' without probing the respondent for a valid response.

however, was on the extent of satisficing in each of the treatment groups as measured by non-differentiation, acquiescence, preference for mid-points and response order effects in rating scales.

To measure each of these four response sets, we used a series of question batteries sharing common response scales. These were:

- 18 agree/disagree items using a five-point rating scale, fully-labelled: *agree strongly; agree; neither agree nor disagree; disagree; disagree strongly*.
- 15 items using a four-point scale labelled: *none or almost none of the time; some of the time; most of the time; all or almost of the time*.
- 5 items using an anchored 7-point scale, where the end-points were labelled *not at all* and *a great deal*[8].

Based on subsets of these items (shown in table 10 below), the respondent's score on each of the indicators of satisficing was calculated by counting the number of times the respondent selected the relevant category and dividing it by the number of items in the set. For acquiescence, for example, the score was calculated as the number of times the respondent selected the agree category. For non-differentiation the score was the maximum number of times they selected *the same response alternative*, regardless of which one it was. These scores were then transformed to run from 0 to 1 for each scale. Table 10 shows the item batteries that were used in each of the four satisficing indicators.

Table 10 – Scales analysed and number of items

Indicator of satisficing	Rating scales analysed	No. of items
Non-differentiation	All scales	38
Acquiescence (agree)	Agree/disagree scales	18
Mid-points	Agree/disagree scales	18
Primacy:		
Strongly agree[9]	Agree/disagree scales	18
Not at all	Not at all/great deal scales	5
Recency:		
Strongly disagree	Agree/disagree scales	18
A great deal	Not at all/great deal scales	5

To compare scores between the groups on each indicator, we initially used t-tests to test the difference in scale means for each of the indicators pairwise between each comparison group (A-B; A-C; B-C). We then ran a series of OLS regression models where we evaluate these group differences with the addition of several covariate main effects and interactions. We present four nested models for each indicator of satisficing in our results. The rationale for the inclusion of covariates in each of the four models follows in the next section.

Model 1

In the baseline model, along with dummy variables for versions B and C we included sex (coded 1 for male) and age (measured as a continuous variable) and education. Although we didn't see statistically significant differences in these variables between groups, there were some small marginal differences and we included them in order to gain precision in our

[8] Question wording for all items is shown in the appendix.

[9] Because the agree-disagree scale and the 'not at all'/'a great deal' scales were of different kinds (5-point, fully-labelled vs. 7-point semantic differential), we decided to treat the two measures of primacy and recency separately, rather than to pool the data from both.

estimates. Education was measured by an item asking respondents to report the total number of years of full-time education they had completed. The distribution of responses to this education measure was similar within each of the countries, so we created a dummy variable that split the sample at the 33rd decile (16 years of education). For all countries this is likely to include those with post-high school education.

Model 2

In the second model, we added country to our baseline model. We included dummies for Hungary, Poland and Switzerland, leaving Germany as the reference category. Our main concern here was to control for unintended differences in survey procedures across countries could account for any effects of questionnaire length on data quality we might observe, as well as to independently examine any variations in response styles across countries.

Model 3

One factor that has been shown in other studies to contribute to satisficing in telephone surveys is the pace at which the interview is conducted (e.g. Holbrook et al., 2003). Pace is of particular interest here because although our experimental treatment manipulated questionnaire length, it was not clear at the outset to what extent it would actually influence interview duration. In general, telephone interviews tend to be conducted at a faster pace compared to face-to-face interviews (e.g. de Leeuw, 1992) – an effect that is assumed to increase the difficulty of the response task. It is not clear however, to what extent respondents feel able or motivated to control the pace at which the interview is conducted, though it seems likely that this will influence the likelihood of satisficing. For example, if respondents are able (and willing) to influence the speed at which the questions are administered, this presumably would facilitate the response task, allowing them to take their time to answer the questions optimally. By contrast, if the interviewer sets the pace too fast and the respondent is unmotivated to slow down, then he or she may be more vulnerable to taking shortcuts in the response process. For these reasons, we decided to include a measure of pace in our third model to control for individual differences in interview length that might independently exert an influence on the likelihood of satisficing. Pace was measured as the total length of the interview in minutes (for group C this was the total length of part 2) divided by the number of items in the questionnaire.

Additionally, we tested for a series of interactions. We included interaction terms for education and the experimental treatment dummies. Our hypothesis was that if satisficing is influenced by the length of the questionnaire, we would expect this effect to be stronger among respondents with less education. Secondly, we included interaction terms for country and the treatment dummies, on the assumption that any effects of questionnaire length on satisficing might vary by country. The inclusion of neither interaction term significantly affected the results so we do not present this analysis here.

Comparing data quality across modes

Having run the above analysis to compare the data quality in the three versions of the telephone experiment, we then repeated parts of it comparing version A of the telephone experiment with the round 3 face-to-face data. In the OLS regression, variables were included that were found to differ significantly across samples, in order to determine whether differences in propensity to satisfice across modes were due to differences in the sample composition or due to mode. Version A was selected because module E appears at the same point (that is, after the same number of items) in version A and in the face-to-face

questionnaire. We would expect to see more evidence of satisficing in the telephone data (see section 3) although the different indicators of satisficing might behave differently by mode.

5.3 Additional analysis

Comparisons of response propensity and data quality by interview length were the primary focus of the analysis of this study and in addition we were able to make some comparisons by mode. We also conducted some supplementary analyses that were useful to enhance our understanding of the differences between telephone and face-to-face interviewing. It was necessary to look at actual length of interview in each group and for each country to ensure that our manipulation of the questionnaire had created the different lengths intended. Similarly, we compared the length of Version A of the telephone questionnaire with the Round 3 face-to-face data so that we could be sure that any differences in data quality were due to mode of administration and not to large differences in length of interview.

Finally, we considered the feedback from survey agencies regarding various elements of the telephone survey.

6 Results

The results are split into two sections, the first looking at the effect of length of interview on rates of cooperation and the second at the effect of length on data quality. For the first part we present the final outcome rates of the telephone survey in each country, comparing results across the three treatment groups to assess the effect of interview length on survey participation. We focus in particular on overall and full rates of cooperation and refusal rates across the three groups, as these most clearly demonstrate the effect of varying estimates of interview length in the survey introduction on respondents' willingness to participate. We then compare the rates of participation between version A of the telephone survey and the round 3 face-to-face survey (both one-hour long). We then look at differences in the socio-demographic composition of the achieved samples in each group to gauge the possible impact of any observed differences in response rates on the likelihood of nonresponse bias in the sample. Finally, we compare the socio-demographic compositions of the achieved samples in version A of the telephone survey and the face-to-face survey in order to identify differential nonresponse bias across mode of data collection.

For the second part, we first describe the composition of the samples in each of the three treatment groups used for the data quality analysis (because the sample used to measure data quality differed slightly from that used to measure response rates. The face-to-face samples used did not differ.). We look at the actual differences in interview length and the pace at which the interviews were conducted, to assess the extent to which interview duration varied between the groups. Next we look at differences in data quality at the item level between the groups by testing the similarity of response distributions and rates of item non-response across groups. We then present the results of our pairwise comparisons and OLS models to assess the effect of questionnaire length on satisficing. Finally, we repeat most of the above analysis of data quality, comparing results between version A of the telephone survey and the round 3 face-to-face data in order to compare data quality across questionnaires of the same length in different modes.

6.1 Rates of participation

6.1.1 Final outcomes: telephone survey

Tables A1 to A5 in Annex 5 show the results for each country on all seven of the outcome indicators described in section 5.1: full and overall response rates (AAPOR RR1 and RR2), full and overall cooperation rates (AAPOR COOP1 and COOP2), household and respondent contact rates (AAPOR CON1) and refusal rates (AAPOR REF1). Response rates, cooperation rates and refusal rates by country and treatment group are shown in table 11. For the purposes of the response rates calculations, the following cases were considered complete: all complete interviews; all complete interviews of version C part 1 only; any near-complete partials (i.e. all the key socio-demographic variables answered).

Table 11 – Response, cooperation and refusal rates by country and treatment group

	A (%)	**B (%)**	**C[10] (%)**	**All (%)**	***n***	**X^2**	**p**
Full Response Rate (RR1)							
Cyprus	8.5	5.3	5.2	6.5	*65*	4.30	.116
Germany	20.3	25.0	21.3	21.6	*329*	2.68	.261
Hungary	18.0	22.0	23.5	21.0	*210*	3.80	.150
Poland	32.1	37.0	25.6	30.4	*292*	8.44	.015*
Switzerland	37.9	39.5	26.8	35.5	*293*	9.45	.009*
Overall Response Rate (RR2)							
Cyprus	8.8	6.2	9.0	8.3	*83*	1.55	.460
Germany	21.8	26.3	25.7	24.3	*369*	3.35	.188
Hungary	19.3	24.5	31.5	25.2	*252*	15.99	.000*
Poland	34.1	38.0	35.3	35.3	*339*	0.86	.652
Switzerland	37.9	39.5	50.2	41.5	*342*	9.05	.011*
Full Cooperation Rate (COOP1)							
Cyprus	13.8	10.3	7.8	10.7	*65*	4.64	.098
Germany	24.7	28.8	25.5	25.9	*329*	1.53	.467
Hungary	19.9	25.6	25.7	23.4	*210*	3.93	.140
Poland	42.4	54.0	36.7	42.3	*292*	10.46	.005*
Switzerland	41.3	42.1	29.0	38.4	*293*	9.66	.008*
Overall Cooperation Rate (COOP2)							
Cyprus	13.1	10.5	12.6	12.4	*83*	0.56	.758
Germany	26.6	30.3	30.8	29.1	*369*	2.49	.288
Hungary	21.3	28.5	34.4	28.0	*252*	15.47	.000*
Poland	45.1	55.6	50.6	49.1	*339*	4.21	.122
Switzerland	41.3	42.1	54.4	44.8	*342*	9.63	.008*

[10] Which parts of version C are included will depend on the different response rate definitions. This will be explained in the text.

	A (%)	B (%)	C (%)	All (%)	*n*	X2	p
Refusal Rate							
Cyprus	30.5	24.9	37.1	31.9	*318*	9.96	.007*
Germany	58.2	58.2	54.2	56.6	*860*	2.41	.300
Hungary	62.3	53.5	53.3	56.9	*569*	7.78	.020*
Poland	23.9	20.1	20.9	43.5	*211*	1.50	.472
Switzerland	31.8	32.4	23.9	29.9	*247*	4.85	.088

Before focusing on the differences between treatment groups within countries, we first consider the results of the survey experiment within each country. As mentioned previously, due to the premature termination of the fieldwork period, response rates were lowest overall in Cyprus. As is evident in table A1 in annex 5, the call record data were incomplete and contained several serious errors affecting around 15% of the sampled cases. For these reasons, the results presented here should be interpreted with caution. They are presented for information only, and we focus the rest of our analysis on the other four countries.

Overall cooperation rates (i.e. the proportion of all those who were willing to be interviewed of all those contacted) were highest in Poland (49%) and Switzerland (45%). Germany and Hungary had comparable rates of overall cooperation at 29% and 28% respectively. In two out of five countries (Hungary and Switzerland), the pattern of results was in the expected direction and in both cases the differences observed were statistically significant at below the 1% level. Cooperation rates were highest in group C, (that is, respondents who were asked to participate in a 30 minute interview). In Poland, the highest rate of cooperation was in group B , though the 30 minute interview still attracted more respondents than the hour-long interview offered to those allocated to group A. In Germany, there was no difference in overall cooperation rates between groups B and C, though both attracted more respondents than the version A questionnaire.

Full cooperation rates (COOP1) discount any partial interviews, which for cases allocated to group C includes all those who refused to respond to the second part of the interview. In both Poland and Switzerland, the proportion of complete interviews in group C was significantly lower than in groups A and B. For example, in Poland, only 37% of those contacted completed the whole of version C of the questionnaire, compared with 54% of those in group B and 42% of those in group A. In Switzerland, the full cooperation rate in group C was 29% compared with 41% in group A and 42% in group B. In Germany and Hungary the differences between the groups were less marked, but cooperation rates were still higher among those in groups B and C compared to group A, though the differences were not statistically significant.

These differences in cooperation were also reflected in variation in refusal rates between the three groups. Note, however, that the refusal rates reported here only reflect refusals made at the last call attempt, and do not include refusals among group C respondents to participate in part 2 of the interview. In all countries (except Cyprus, though for the reasons explained we do not consider the results to be valid), the pattern of refusals was in the direction expected, with higher rates of refusals for longer questionnaires. The differences were not large, however, and the size of the relative differences between the pairs of groups was not consistent. In Germany and Switzerland, there was no difference in refusal rates between groups A and B, but refusal rates were lower in group C by comparison (though the differences were not significant). In Hungary and Poland, there was no difference between the 45 minute and the 30 minute interview (groups B and C), whereas both were less likely to

result in a refusal than the invitation to participate in a 60 minute interview (group A). In Hungary the difference was statistically significant.

Similar patterns of findings were evident in overall response rates, with no differences between groups B and C in Germany and Poland, but small (non-significant) differences between both groups and group A. By contrast, the overall response rates were significantly higher in group C compared with groups A and B in both Hungary and Switzerland. Removing the partial interviews and respondents in group C who only completed part 1 of the interview led to significantly lower full response rates in groups C compared with A and B in both Poland and Switzerland. In Germany and Hungary, however, this only had the effect of bringing the response rates in group C more in line with those in group B (though in both groups, response rates were higher than in group A).

6.1.2 Final outcomes: comparison of telephone and face-to-face

***Table 12 – Comparison of response rates: telephone experiment and round 3 face-to-face survey*[1]**

	Full response rate (ESS response rate)		**Full co-operation rate**		**Refusal rate**		**Non-contact rate**	
	Face-to-face	**Telephone version A**	**Face-to-face**	**Telephone version A**	**Face-to-face**	**Telephone version A**	**Face-to-face**	**Telephone version A**
Cyprus[2]	67.3%	8.5%	93.4%	13.8%	4.2%	30.5%	27.9%	57.8%
Germany	54.5%	20.3%	57.4%	24.7%	23.7%	58.2%	5.1%	17.9%
Hungary	66.1%	18%	69.7%	19.9%	25.2%	62.3%	5.2%	9.8%
Poland	70.2%	32.1%	74.8%	42.4%	15.8%	23.9%	6.2%	24.4%
Switzerland[3]	45.8%	37.9%	50.8%	41.3%	39.2%	31.8%	8.9%	9.4%

Notes: [1]Figures taken from 'ESS3-2006 Documentation Report, ed. 3.2' based on information provided by the survey agencies [2]In Cyprus a large number of selected units (381 out of 1481) were dropped when the target number of interviews was achieved. These were counted as non-contacts. 3 The Swiss face-to-face data are only from the two French-speaking regions, as with the telephone data.

Table 12 above shows a comparison of rates of participation in the telephone experiment and the round 3 face-to-face survey. Only version A has been included in this comparison so that we are only comparing the rates of participation for the hour-long interviews. This table does not consider the overall response and co-operation rates (that is the proportion of individuals who agreed to take part, regardless of whether or not they went on to complete the questionnaire) because comparable data are not available for the face-to-face and telephone data. Data were not available for the number of partial cases and, as a result, it is not possible to calculate the 'overall' rates of response and co-operation for the face-to-face survey. So in the table above, and the discussion below, we concentrate on the full response and cooperation rates, the refusal rate and the non-contact rate.

The different sample designs for the telephone survey should be considered when interpreting these results (see section 4.2). For example, for telephone surveys, in particular those using RDD, it can be difficult to distinguish a non-contact from an ineligible case. Whereas an interviewer visiting an address can usually tell if the address is, or is not, a residential household, this cannot always be determined over the phone. Interviewers were instructed to

only classify a case as 'ineligible' if there was no doubt that it was ineligible. Otherwise it should be coded 'non-contact'.

The figures for Cyprus are difficult to interpret because there were problems in both the face-to-face and telephone survey (see page 8 of the 'ESS3 Response Based Quality Assessment' report for details of the face-to-face fieldwork: http://ess.nsd.uib.no/ess/round3/surveydoc.html). They will not be discussed here.

In all countries the face-to-face survey response rate is higher than that for the telephone survey. In Poland, Hungary and Germany the differences between the face-to-face and telephone response rates range between 34.2 and 48.1 percentage points. In Hungary and Germany this seems to be mostly down to higher refusal rates rather than to non-contact rates, which suggest that respondents can be contacted by phone but are unwilling to take part in an hour-long survey over the telephone. In Hungary, the non-contact rates are more similar between modes than in the other countries (5.2% face-to-face vs 9.8% telephone).

The difference in refusal rate is quite low in Poland, with the telephone experiment getting just 8.1% more refusals than the face-to-face survey. In fact, the refusal rate for the telephone survey in Poland (23.9%) was as low as, or lower than, the face-to-face refusal rates in Germany, Hungary and Switzerland. However, although Poland achieved the highest co-operation rate (42.4%), the co-operation rate is still much lower than that achieved in the face-to-face survey in Poland (74.8%).

Response rates between modes were most similar in Switzerland – the full response rate in the face-to-face survey (45.8%) is just 7.9 percentage points higher than in the telephone survey (37.9%). Furthermore, it should be noted that in the face-to-face survey in Switzerland a great deal of effort and expense is put into increasing the response rate. In round 1 the response rate was 33.5%, 4.4 percentage points *lower* than the response rate achieved in the telephone experiment.

Much anecdotal evidence suggests that people in some countries are more used to being contacted by telephone than face-to-face and as a result are more likely to refuse when approached in person. It has been suggested that this is the case in Switzerland and the evidence from this experiment seems to support that. The refusal rate for version A of the telephone experiment in Switzerland (31.8%) was 7.4 percentage points *lower* than the refusal rate in the face-to-face survey (39.2%). Although the non-contact rate is higher for the telephone experiment, it is only by 6.4 percentage points (although, note that the non-contact rate for Switzerland as a whole was lower than that for the French –speaking regions only).

6.1.3 Sample composition: telephone survey

Next we consider the extent to which the observed differences in response rates resulted in differences in the composition of the achieved samples in each group. Group C respondents who only completed the first part of the questionnaire were included in this analysis, as were respondents who completed the survey over multiple occasions. Table 13 shows the number of cases analysed in each of the three groups of interest.

Table 13 – Cases analysed by treatment group and country

	Group A	Group B[1]	Group C	Total
Germany	123	76	177	376
Hungary	72	45	122	239
Poland	131	71	145	347
Switzerland	154	83	105	342
Total	480	275	549	1304

Notes: [1] The issued sample size for group B was half that for groups B and C.

Tables 14-17 show the socio-demographic makeup of the samples across each of the three treatment groups in each country. We look here at sex and age of respondents in each group, whether or not they were in paid work at the time of the interview, what kind of area they live in, their main activity, and number of years of education (all variables that have been found in other studies to be associated with response propensities in surveys). For group C we separate respondents who responded to both parts of the interview from those responding only to part 1.

There was little evidence of differential nonresponse between the treatment groups. Few differences were observed between the samples on the socio-demographic variables considered here. In Germany and Switzerland there were no significant differences between groups. In Hungary, respondents in group C (both parts) were significantly more like to have been in paid work at the time of the interview. There was also a significantly lower proportion of male respondents in group C (both parts) in Poland.

Table 14 – Socio-demographic composition by interview length - Germany

Version	A	B	C (both parts)	C (part 1 only)	Total	Test-statistic	P-value
n	123	76	130	47	376		
Male %	53.7	44.7	56.2	42.6	51.3	4.25 (X^2)	.236
Mean age (years)	48.7	47.9	49.7	49.4	49.0	.18 (F)	.911
Currently in paid work (%)	58.2	55.3	56.2	55.3	56.5	.22 (X^2)	.974
Area						9.39 (X^2)	.311
Big city	14.6	15.8	10.0		13.1		
Suburb	14.6	15.8	23.1		18.2		
Town	37.4	32.9	40.8		37.7		
Village	28.5	34.2	23.8		28.0		
Farm	4.9	1.3	2.3		3.0		
Main activity						27.18 (X^2)	.165
Paid work	52.0	46.1	45.4	48.9	48.1		
Education	12.2	9.2	13.1	8.5	11.4		
Unemployed, looking	2.4	0.0	4.6	4.3	2.9		
Unemployed, not looking	0.8	0.0	1.5	2.1	1.1		
Permanently sick or disabled	0.0	5.3	2.3	0.0	1.9		
Retired	25.2	23.7	30.0	25.5	26.6		
Housework, caring for children	3.3	7.9	2.3	6.4	4.3		
Other	4.1	7.9	0.8	4.3	3.7		
Education years (mean)	14.6	14.5	14.8	15.1	14.7	.34 (F)	.796

Table 15 – Socio-demographic composition by interview length- Hungary

Version	A	B	C (both parts)	C (part 1 only)	Total	Test-statistic	P-value
n	72	45	94	28	239		
Male %	37.5	40.0	25.5	32.1	32.6	4.05(X^2)	.257
Mean age (years)	49.7	49.9	52.5	52.9	51.2	.57(F)	.635
Currently in paid work (%)	33.3	46.7	55.3	42.9	45.6	8.05(X^2)	.045*
Area						7.47(X^2)	.486
Big city	18.1	13.3	12.8		14.7		
Suburb	9.7	17.8	7.4		10.4		
Town	38.9	31.1	38.3		37.0		
Village	31.9	37.8	41.5		37.4		
Farm	1.4	0.0	0.0		0.5		
Main activity						42.07(X^2)	.004**
Paid work	31.0	40.0	48.9	35.7	40.3		
Education	16.9	6.7	2.1	3.6	7.6		
Unemployed, looking	4.2	0.0	1.1	0.0	1.7		
Unemployed, not looking	0.0	0.0	1.1	0.0	0.4		
Permanently sick or disabled	8.5	6.7	2.1	0.0	4.6		
Retired	28.2	35.6	42.6	50.0	37.8		
Housework, caring for children	9.9	8.9	2.1	3.6	5.9		
Other	1.4	2.2	0.0	7.1	1.7		
Education years (mean)	13.9	14.5	14.2	14.4	14.2	.23 (F)	.875

Table 16 – Socio-demographic composition by interview length- Poland

Version	A	B	C (both parts)	C (part 1 only)	Total	Test-statistic	P-value
n	131	71	100	45	347		
Male %	44.3	43.7	28	48.9	40.1	8.87 (X^2)	.031*
Mean age (years)	45.4	49.2	50.1	46.6	47.7	2.02 (F)	.111
Currently in paid work (%)	51.9	56.3	44.0	55.6	51.0	3.19 (X^2)	.364
Area						3.12 (X^2)	.927
Big city	15.3	16.9	18.0		16.6		
Suburb	9.9	5.6	12.0		9.6		
Town	34.4	31.0	30.0		32.1		
Village	32.1	36.6	33.0		33.4		
Farm	8.4	9.9	7.0		8.3		
Main activity						27.55 (X^2)	.153
Paid work	47.3	50.7	36.0	51.1	45.2		
Education	9.2	2.8	5.0	8.9	6.6		
Unemployed, looking	0.0	2.8	0.0	0.0	0.6		
Unemployed, not looking	0.8	1.4	1.0	4.4	1.4		
Sick or disabled	0.0	0.0	1.0	0.0	0.3		
Retired	28.2	29.6	42.0	28.9	32.6		
Housework, caring for children	9.9	9.9	13.0	4.4	10.1		
Other	4.6	2.8	2.0	2.2	3.2		
Education years (mean)	13.03	13.18	13.29	12.20	13.03	1.28 (F)	.281

Table 17 – Socio-demographic composition by interview length - Switzerland

Version	A	B	C (both parts)	C (part 1 only)	Total	Test-statistic	P-value
n	154	83	56	49	342		
Male %	44.2	47.0	42.9	44.9	44.7	.27 (X^2)	.965
Mean age (years)	47.2	45.4	47.3	48.2	46.9	.29 (F)	.831
Currently in paid work (%)	33.8	36.1	30.4	34.7	33.9	.52 (X^2)	.916
Area						5.00 (X^2)	.758
Big city	16.2	19.3	21.4		18.1		
Suburb	9.1	10.8	7.1		9.2		
Town	27.3	27.7	17.9		25.6		
Village	39.6	38.6	44.6		40.3		
Farm	7.8	3.6	8.9		6.8		
Main activity						24.52 (X^2)	.269
Paid work	59.7	56.6	58.9	57.1	58.5		
Education	6.5	10.8	7.1	6.1	7.6		
Unemployed, looking	0.0	0.0	0.0	2.0	0.3		
Unemployed, not looking	0.6	0.0	0.0	2.0	0.6		
Sick or disabled	1.3	0.0	7.1	4.1	2.3		
Retired	21.4	22.9	17.9	22.4	21.3		
Housework, caring for children	9.1	6.0	5.4	2.0	6.7		
Other	1.3	3.6	3.6	4.1	2.6		
Education years (mean)	14.2	13.9	15.3	13.7	14.3	1.81 (F)	.146

6.1.4 Sample composition: comparison of face-to-face and telephone

In order to find out whether there were any differences in the sample composition of the telephone and face-to-face samples, that is to see whether different types of people respond to different modes, we compared the sample on a number of socio-demographic variables: gender; age; whether or not the respondent is currently in paid work; the type of area in which the respondent lives; their main activity; and the number of years they were in education.

All analysis presented here used unweighted data. Only data for French-speaking respondents is included in the telephone and face-to-face sample for Switzerland.

Table 18 - Sample sizes

	Face-to-face	Telephone	Total
Poland	1218	99	1317
Switzerland	358	131	489
Hungary	855	69	924
Germany	2663	123	2786
Total	6512	422	6934

Table 19 - Socio-demographic composition by mode - all countries

	Face-to-face	Telephone	Total	Test statistic	P-Value
Male (%)	**46.0**	**45.5**	**46.0**	**.045 (X^2)**	**.832**
Mean age (years)	**48.7**	**47.6**	**48.6**	**1.188 (t)**	**.235**
Currently in paid work (%)	**51.2**	**44.2**	**50.7**	**7.727 (X^2)**	**.005****
Area (%)				**11.283 (X^2)**	**.010****
Big city	**20.2**	**15.2**	**19.9**		
Suburb	**11.1**	**11.6**	**11.1**		
Town	**36.6**	**34.4**	**36.5**		
Village / farm	**32.0**	**38.9**	**32.6**		
Main activity (%)				**6.990 (X^2)**	**.136**
Paid work	**45.7**	**50.6**	**46.1**		
Education	**9.2**	**10.5**	**9.3**		
Retired	**27.7**	**24.9**	**27.5**		
Housework, caring for children	**10.1**	**7.4**	**9.9**		
Other	**7.3**	**6.7**	**7.2**		
Education years (mean)	**12.8**	**14.0**	**12.9**	**-6.519 (t)**	**.000*****

Note to tables 19-23: Income is not included in these tables as it is in the equivalent tables for the telephone comparisons (tables 14-17). Because income was asked differently in telephone and face-to-face administration we do not have comparable data to create the same income variable that was used for the telephone analysis**.**

Table 20 - Socio-demographic composition my mode - Germany

	Face-to-face	Telephone	Total	Test statistic	P-Value
Male (%)	**48.6**	**53.7**	**48.8**	**1.208 (X^2)**	**.272**
Mean age (years)	**48.4**	**48.7**	**48.4**	**-.203 (t)**	**.839**
Currently in paid work (%)	**53.3**	**41.8**	**52.8**	**6.176 (X^2)**	**.013****
Area (%)				**2.647 (X^2)**	**.449**
Big city	**17.1**	**14.6**	**17.0**		
Suburb	**15.8**	**14.6**	**15.8**		
Town	**40.2**	**37.4**	**40.1**		
Village / farm	**26.8**	**33.3**	**27.1**		
Main activity (%)				**9.073 (X^2)**	**.059**
Paid work	**47.3**	**52.0**	**47.5**		
Education	**9.0**	**12.2**	**9.1**		
Retired	**24.1**	**25.2**	**24.1**		
Housework, caring for children	**11.3**	**3.3**	**11.0**		
Other	**8.3**	**7.3**	**8.3**		
Education years (mean)	**13.3**	**14.6**	**13.3**	**-4.156(t)**	**.000*****

Table 21 - Socio-demographic composition my mode - Hungary

	Face-to-face	Telephone	Total	Test statistic	P-Value
Male (%)	**38.7**	**36.2**	**38.5**	**.166 (X^2)**	**.684**
Mean age (years)	**55.0**	**49.4**	**54.6**	**2.423 (t)**	**.016***
Currently in paid work (%)	**41.9**	**65.2**	**43.6**	**14.172 (X^2)**	**.000*****
Area (%)				**7.638 (X^2)**	**.054***
Big city	**23.6**	**15.9**	**23.1**		
Suburb	**4.0**	**10.1**	**4.4**		
Town	**36.0**	**40.6**	**36.4**		
Village / farm	**36.4**	**33.3**	**36.1**		
Main activity (%)				**25.527 (X^2)**	**.000*****
Paid work	**38.3**	**32.4**	**37.9**		
Education	**4.9**	**16.2**	**5.8**		
Retired	**42.9**	**26.5**	**41.7**		
Housework, caring for children	**7.5**	**10.3**	**7.7**		
Other	**6.3**	**14.7**	**6.9**		
Education years (mean)	**12.3**	**13.8**	**12.4**	**-2.972 (t)**	**.003****

Table 22 - Socio-demographic composition my mode - Poland

	Face-to-face	Telephone	Total	Test statistic	P-Value
Male (%)	**46.7**	**43.4**	**46.5**	**.396 (X^2)**	**.529**
Mean age (years)	**44.9**	**45.8**	**44.9**	**-.548 (t)**	**.584**
Currently in paid work (%)	**50.1**	**46.5**	**49.8**	**.490 (X^2)**	**.484**
Area (%)				**13.179 (X^2)**	**.004****
Big city	**27.3**	**16.2**	**26.4**		
Suburb	**4.9**	**12.1**	**5.5**		
Town	**32.6**	**35.4**	**32.8**		
Village / farm	**35.2**	**36.4**	**35.3**		
Main activity (%)				**2.211 (X^2)**	**.697**
Paid work	**45.2**	**49.5**	**45.6**		
Education	**13.7**	**9.1**	**13.4**		
Retired	**27.2**	**29.3**	**27.3**		
Housework, caring for children	**7.7**	**7.1**	**7.7**		
Other	**6.2**	**5.1**	**6.1**		
Education years (mean)	**11.7**	**13.2**	**11.8**	**-4.197 (t)**	**.000*****

Table 23 - Socio-demographic composition my mode - Switzerland

	Face-to-face	Telephone	Total	Test statistic	P-Value
Male (%)	**42.2**	**44.3**	**42.7**	**.172 (X^2)**	**.678**
Mean age (years)	**48.9**	**47.0**	**48.4**	**1.027 (t)**	**.305**
Currently in paid work (%)	**62.0**	**33.6**	**54.4**	**31.234 (X^2)**	**.000*****
Area (%)				**2.912 (X^2)**	**.405**
Big city	**11.2**	**14.5**	**12.1**		
Suburb	**14.0**	**9.2**	**12.7**		
Town	**24.9**	**27.5**	**25.6**		
Village / farm	**50.0**	**48.9**	**49.7**		
Main activity (%)				**4.791 (X^2)**	**.309**
Paid work	**53.0**	**59.5**	**54.7**		
Education	**5.4**	**6.9**	**5.8**		
Retired	**20.3**	**20.6**	**20.4**		
Housework, caring for children	**15.5**	**9.9**	**14.0**		
Other	**5.9**	**3.1**	**5.1**		
Education years (mean)	**13.8**	**14.1**	**13.9**	**-.778 (t)**	**.437**

For most of the socio-demographic variables we tested, there was no significant difference between the sample obtained in the face-to-face survey and that of the telephone survey. However, there were some significant differences and two variables (years in education and being in paid work) differed significantly by mode in almost all countries (with the exception of paid work in Poland and years in education in Switzerland).

Number of years in education is significantly higher in the telephone sample in all countries except Switzerland. This suggests that more highly educated people are more likely to co-operate in the survey by telephone, either because they can be more easily contacted by that mode, or more likely to be persuaded to participate.

In all countries, except for Hungary, the proportion of respondents in paid work is lower in the telephone sample (significant in Switzerland, Germany and for all countries combined), suggesting perhaps that people who work are easier to contact and persuade to participate by telephone than in person.

In Poland the type of area in which the respondent lives differs significantly by mode and it is approaching significance in Hungary. In both countries respondents in the telephone sample are less likely to live in big cities (this is also true in Germany but the results are not significant). In Switzerland, area is not significant.

In Hungary, respondent's age, number of years in education and main activity are also significant, with those in the telephone sample being younger, more educated, more likely to currently be in education and less likely to be retired.

6.2 Data quality

6.2.1 Effect of length of interview on data quality

In order to explore variation in data quality as a function of questionnaire length, we analysed data from cases with complete interviews in each of the experimental groups. For group C we retained those respondents who elected to schedule a new appointment for the part 2 interview but removed those who chose to continue with the second part of the interview straight after

the first (around 30% of group C respondents in total)[11] to ensure consistency in the number of items preceding module E. This reduced the number of cases available for analysis, which led to the decision to pool data from all four countries. Although we were interested to see whether data quality might also vary cross-nationally[12], we had no theoretical grounds to assume that the *predictors* of satisficing would be different in each country, so we felt justified in combining the data in this way. However, in order to control for any possible cross-national differences in the effects we were interested in observing, we included country controls in our OLS regression models (see below). Table 24 shows the number of cases analysed in each of the three groups of interest.

Table 24 – Cases analysed by treatment group and country

	Group A	Group B[1]	Group C[2]	Total
Germany	123	76	107	306
Hungary	69	44	42	155
Poland	99	58	75	232
Switzerland	131	80	42	253
Total	422	258	266	946

Notes: [1] The issued sample size for group B was half that for groups B and C. [2] Group C includes only those respondents who were interviewed on two separate occasions.

Before comparing data quality across the three groups, the samples in each country were compared on a range of socio-demographics to verify that there were no differences in sample composition that might account for any observed differences in response. Although this had already been done for the response rate analysis data (see section 6.1.3), it was repeated because the sample used differed slightly. For this section of the analysis, we analysed all complete (or near complete) interviews conducted on one occasion (for groups A and B) and on two occasions (for group C). No significant differences were found, though one or two were approaching significance. For simplicity, and because no significant results were found on the observed variables we examined here, table 25 shows this information pooled for all countries combined. The difference in mean age is approaching significance, with a higher proportion of older people in group C (the shortest questionnaire according to the introduction). Women were over-represented in each group, and slightly more so in group C, though this difference was not statistically significant.

[11] As a result the telephone samples used for the analysis of response rates and data quality differ slightly. The same face-to-face sample was used for all analyses.

[12] Several studies have found evidence of cultural differences in response effects such as extreme response style and acquiescence (Hui and Triandis, 1989; Clarke, 2001) and social desirability bias (Johnson and van der Vijver, 2003).

Table 25 – Socio-demographic composition of the sample. All countries

Version	A	B	C	Total	Test statistic	P-Value
n	**422**	**258**	**266**	**946**		
Male (%)	**45.5**	**45.0**	**40.2**	**43.9**	**1.98 (X^2)**	**.371**
Mean age (years)	**47.6**	**47.7**	**50.7**	**48.5**	**2.85 (F)**	**.058**
Currently in paid work (%)	**44.2**	**42.6**	**44.8**	**43.9**	**0.27 (X^2)**	**.873**
Area (%)					**5.20 (X^2)**	**.736**
Big city	15.2	16.3	13.8	15.1		
Suburb	11.6	12.8	14.9	12.9		
Town	34.4	31.0	35.2	33.7		
Village	32.9	36.0	31.8	33.5		
Farm	5.9	3.9	4.2	4.9		
Main activity (%)					**20.65 (X^2)**	**.111**
Paid work	50.6	50.0	46.4	49.3		
Education	10.5	7.8	6.9	8.7		
Unemployed, looking	1.4	0.4	1.9	1.3		
Unemployed, not looking	0.7	0.4	1.5	0.9		
Permanently sick or disabled	1.9	2.7	2.7	2.3		
Retired	24.9	26.7	34.5	28.1		
Housework, caring for children	7.4	7.8	5.0	6.8		
Other	2.6	4.3	1.1	2.7		
High income (%)[1]	**32.6**	**26.5**	**25.2**	**28.8**	**4.36 (X^2)**	**.113**
Education years (mean)	**14.0**	**14.0**	**14.3**	**14.1**	**0.62 (F)**	**.538**

Notes: [1]'High income' was calculated by combining the top 3rd income group for each country.

As a reminder, while our treatment groups varied in terms of questionnaire length, what we were interested in here was the length (or number of items) preceding module E. In fact, the versions are in the same order for length of whole interview and length before module E. Version A had the most items before module E (141), version B the second most (86), and version C the least (36) (where the second part was done on a different occasion to the first part, i.e. as it was for those cases included here).

To establish what the actual differences in *interview* length were between the groups, we looked at interview duration (interviewers were instructed to record the start time and end time of each interview). Table 26 shows the mean interview lengths for the countries as a whole and for each country separately. These were fairly consistent across country and in line with what we had anticipated. The mean scores, however, masked wide variations within the treatment groups in the actual lengths of the interviews. Figure 1 illustrates the range of values for interview duration by treatment group (for group C, part 1 and 2 interview lengths have been combined). Given this variation within groups, we also looked at mean interview pace across each of the groups (calculated as the length of the interview in minutes divided by the number of items in the questionnaire[13]) to assess the extent to which actual interview duration was independent of the experimental treatment. The resulting scores represent the amount of time taken per item. The mean pace of interviews was similar across all three groups (0.23 minutes per question in A, 0.22 in B and 0.21 in C), but the differences between group C and both A and B was statistically significant ($F_{2,943} = 8.74$; $p<0.001$ for A & C and $p<0.05$ for B & C).

[13] For group C, pace was calculated as the length (in minutes) of interview part 2 (in which the module of questions on wellbeing was located) divided by the number of items in part 2.

Figure 1 – Variation in interview length by treatment group

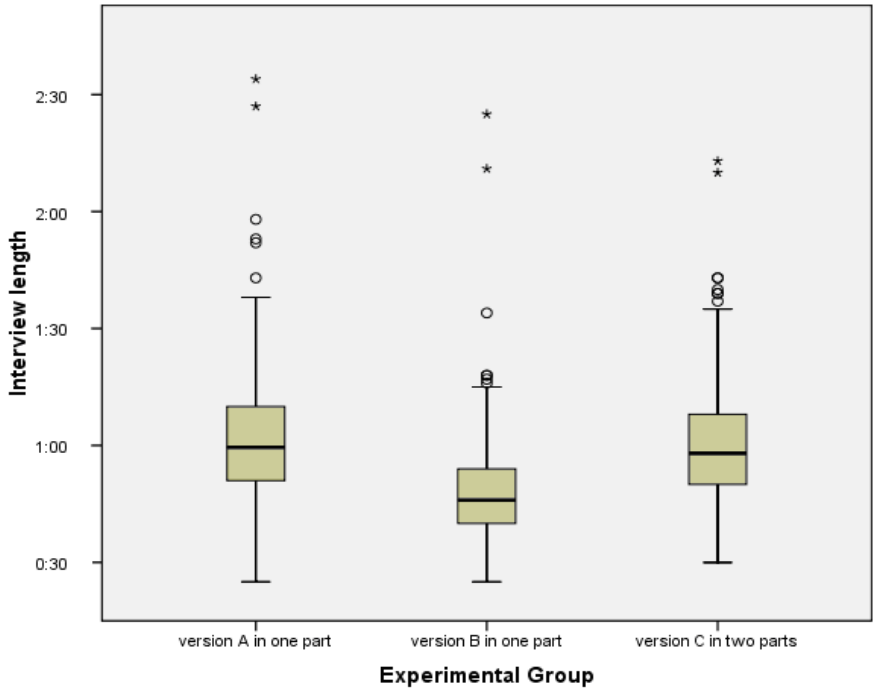

Table 26 – Interview lengths by treatment group and country (hours:minutes)

		N	Min.	Max.	Mean	S.D.
All	All groups	1052	0.25	2.34	0.58	0.15
	Group A	422	0.25	2.34	1.01	0.15
	Group B	258	0.25	2.25	0.48	0.13
	Group C part 1	265	0.09	1.17	0.30	0.11
	Group C part 2	266	0.13	1.44	0.30	0.09
	Group C (all)	265	0.30	2.13	1.00	0.15
Germany	All groups	329	0.25	2.25	0.54	0.13
	Group A	123	0.25	1.35	0.56	0.10
	Group B	76	0.25	2.25	0.43	0.13
	Group C part 1	107	0.19	0.55	0.32	0.07
	Group C part 2	107	0.16	0.50	0.27	0.05
	Group C (all)	107	0.39	1.30	1.00	0.10
Hungary	All groups	203	0.28	2.13	0.57	0.15
	Group A	69	0.37	1.37	0.58	0.12
	Group B	44	0.31	1.09	0.44	0.08
	Group C part 1	42	0.23	0.59	0.35	0.08
	Group C part 2	42	0.13	1.44	0.32	0.14
	Group C (all)	42	0.43	2.13	1.08	0.16
Poland	All groups	263	0.30	2.34	1.01	0.16
	Group A	99	0.42	2.24	1.10	0.15
	Group B	58	0.33	1.18	0.55	0.10
	Group C part 1	74	0.09	1.17	0.21	0.13
	Group C part 2	75[1]	0.20	0.58	0.33	0.08
	Group C (all)	74	0.30	2.10	0.54	0.17

Switzerland	All groups	267	0.31	2.27	0.59	0.16
	Group A	131	0.39	2.27	1.02	0.16
	Group B	80	0.31	2.11	0.50	0.14
	Group C part 1	42	0.20	1.13	0.32	0.09
	Group C part 2	42	0.21	0.54	0.31	0.07
	Group C (all)	42	0.45	1.40	1.03	0.14

Notes: [1] 1 case had a missing value for end time of interview.

As a matter of interest, we also compared the interview lengths for version A of the telephone survey and the face-to-face survey to see how length of interview differs by mode of data collection. Both questionnaires contained the same number of items (with the exception of 5 items added to the telephone questionnaire). There were a number of outliers in the face-to-face data and as a result any cases lasting more than 3 hours and less than 25 minutes were removed (16 cases in total). Based on evidence from other studies, we would expect the telephone interviews to be shorter than the face-to-face ones because they tend to be conducted at a faster pace (Holbrook, Green and Krosnick, 2003), and that is indeed what we found here.

Table 27 – Interview lengths by mode and country (minutes)

		N	Min.	Max.	Mean	S.D.	Sig.
All	Telephone (group A)	422	25	154	62	15.102	***
	Face-to-face	5062	25	180	66	18.473	
Germany	Telephone (group A)	123	25	95	56	10.772	***
	Face-to-face	2650	25	180	65	18.024	
Hungary	Telephone (group A)	69	37	97	59	12.693	***
	Face-to-face	855	32	169	67	16.516	
Poland	Telephone (group A)	99	42	154	70	15.099	*
	Face-to-face	1205	30	180	66	19.851	
Switzerland	Telephone (group A)	131	39	147	62	16.880	***
	Face-to-face	352	36	170	70	20.739	

Notes: * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table 27 shows that face-to-face is longer on average in all countries except Poland where the average length for the telephone interviews is 4 minutes longer than the face-to-face interviews. In all countries except Poland these results are significant at $p<0.001$. However, the substantive differences are not large and the average length in all countries in both modes is around 1 hour. We can also see that the maximum length is higher in the face-to-face interviews, even when outliers have been removed. The biggest difference between modes is in Hungary and Germany where the maximum length differs by 72 and 85 minutes respectively and the average length differs by 8 and 9 minutes respectively.

Turning to our analysis of data quality, the table in the annex 1 shows the results of chi-square tests comparing the marginal distribution of responses to all items in the well-being module (module E). In total, for 11 out of the 46 items (24%) the value for chi-square was statistically significant, indicating that response distributions across the treatment groups were not comparable. Closer inspection of the cross-tabulations indicated that in almost all cases

responses in the short questionnaire group (C) were differently distributed to those in the longer questionnaire groups (A and B). We also examined item non-response rates across the 46 items, but the overall rates of missing values were very low. Refusals and 'No answers' affected only 1 or 2 questions. The mean rate of Don't Know responses (i.e. the total number of 'Don't Know' responses divided by the number of items in the module) was also very low at 0.25 in group A, 0.26 in group B and just 0.18 in group C. If Don't Know reporting is indicative of reduced data quality, then this difference is in the expected direction – i.e. respondents were less likely to answer 'Don't Know' in the short questionnaire group – but it was not statistically significant.

Turning to our hypothesis tests relating to the four indicators of satisficing (see section 5.2), we first ran ANOVA tests to compare mean scores on each one across different pairs of treatment groups in order to get an initial idea of the extent and direction of differences as a function of questionnaire length. Mean scores are shown in table 28.

Table 28 – Mean scores on satisficing indicators – variations in length

Indicator	Mean			Sig.		
	A	B	C[1]	A/C	A/B	B/C
Non-differentiation	0.37	0.36	0.37			
Use of mid-point	0.18	0.19	0.22	***		*
Acquiescence (agree)	0.40	0.40	0.43	*		*
Primacy						
Agree strongly	0.21	0.21	0.16	**		**
Not at all	0.40	0.36	0.30	*		
Recency						
Disagree strongly	0.08	0.07	0.06	**		*
A great deal	0.83	0.75	0.64	*		

Notes: * p<0.05; ** p<0.01; *** p<0.001. [1]Group C here is those who completed both parts of version C but on different occasions.

There were no significant differences between the mean scores in groups A and B. However, there were significant differences between both groups A and C, and groups B and C, though the direction of the differences was mixed. The mean scores for *non-differentiation* were not significantly different in any comparison group. There was significantly more use of mid-points and more acquiescence with the shortest questionnaire (version C), while primacy and recency effects were more prevalent in the longer questionnaires.

These differences were analysed further in our four OLS models (see section 5.2), the results of which are shown in table 29. In model 1 (shown in column 2 of table 29), the dummy for the short questionnaire (version C) was significant on all indicators other than *non-differentiation*. As we saw from the comparison of means analysis, selection of the end-points (primacy and recency) was more common in the long questionnaire while use of mid-points and acquiescence were more common in the short questionnaire. As expected, there was evidence of less satisficing in the high education group, though this effect was not observed on all indicators (only on non-differentiation, acquiescence and primacy on the 7-point semantic differential scale).

Table 29 – Regression coefficients from OLS models

Indicator	Model 1	Model 2	Model 3
Non-differentiation			
- **Constant**	.360	.356	.362
- **Sex**	.008	.009	.009
- **Age**	.000	.000	.000*
- **B (Med)**	-.003	-.002	-.002
- **C (Short)**	.002	-.003	-.004
- **Highed**	-.041***	-.036***	-.036***
- **Switzerland**		-.023**	-.022**
- **Hungary**		-.015	-.015
- **Poland**		.051***	.053***
- **Pace**			-.042
- **Pace*age**			
Acquiescence			
- **Constant**	.395	.412	.403
- **Sex**	.002	-.002	-.002
- **Age**	.000	.000	.000
- **B (Med)**	.002	.001	.002
- **C (Short)**	.032*	.018	.019
- **Highed**	-.044***	-.034**	-.034**
- **Switzerland**		-.058***	-.060***
- **Hungary**		-.086***	-.087***
- **Poland**		.062***	.059***
- **Pace**			.055
- **Pace*age**			
Mid-points			
- **Constant**	.258	.291	.302
- **Sex**	-.018*	-.015	-.015
- **Age**	-.001***	-.002***	-.002***
- **B (Med)**	.010	.015	.015
- **C (Short)**	.043***	.036**	.035**
- **Highed**	-.001	-.003	-.003
- **Switzerland**		-.090***	-.088***
- **Hungary**		.031*	.032*
- **Poland**		-.024*	-.020
- **Pace**			-.069
- **Pace*age**			
Primacy			
Agree strongly			
- **Constant**	.148	.117	.121
- **Sex**	.019	.021	.021
- **Age**	.001**	.001**	.001**
- **B (Med)**	-.001	-.003	-.003
- **C (Short)**	-.047**	-.032*	-.032*
- **Highed**	.020	.014	.014
- **Switzerland**		.099***	.100***
- **Hungary**		.046**	.047**
- **Poland**		-.037*	-.036*
- **Pace**			-.024
- **Pace*age**			

Indicator	Model 1	Model 2	Model 3
Not at all			
- **Constant**	.127	-.091	-.119
- **Sex**	-.090*	-.035	-.036
- **Age**	.008***	.007***	.007***
- **B (Med)**	-.062	-.042	-.042
- **C (Short)**	-.139**	-.085*	-.082
- **Highed**	-.089*	-.099**	-.099**
- **Switzerland**		.222***	.217***
- **Hungary**		.550***	.548***
- **Poland**		.161**	.153**
- **Pace**			.172
- **Pace*age**			
Recency			
Disagree strongly			
- **Constant**	.050	.021	.029
- **Sex**	.004	.005	.005
- **Age**	.001**	.001***	.001***
- **B (Med)**	-.005	-.008	-.008
- **C (Short)**	-.024**	-.014*	-.015*
- **Highed**	.008	.006	.006
- **Switzerland**		.079***	.080***
- **Hungary**		.020*	.020**
- **Poland**		-.005	-.002
- **Pace**			-.050
- **Pace*age**			
A great deal			
- **Constant**	.538	.324	.310
- **Sex**	-.150*	-.064	-.064
- **Age**	.008***	.007***	.007***
- **B (Med)**	-.138	-.098	-.098
- **C (Short)**	-.286**	-.207**	-.205*
- **Highed**	.080	.037	.037
- **Switzerland**		.151	.149
- **Hungary**		.961***	.960***
- **Poland**		-.013	-.017
- **Pace**			.087
- **Pace*age**			

Notes: * p<0.05; ** p<0.01; *** p<0.001.

Turning to model 2, with just a few exceptions, all the country dummies were significant on all indicators, indicating a greater or lesser degree of satisficing in those countries compared with in the German sample (the reference group). However, the direction of the effect of country was not consistent across the indicators. This is perhaps not surprising because, although we would not necessarily expect one country to satisfice more than another, we might expect respondents in different countries to use response options differently (due to cultural reasons, translation, variation in survey practices, etc.). Adding country to the model also had the effect of making sex non-significant on all indicators where it had been significant in model 1, showing that some of the apparent effect of being male was due to compositional differences in the national samples. Notably, the effect of questionnaire length became weaker but remained significant on all indicators where it had been in model 1, with the exception of acquiescence. This seems likely to indicate some differences in either propensity to satisfice across countries, or differences in experimental protocols or recruitment.

In model 3, interview pace was not found to be a significant predictor of satisficing on any of the indicators and adding pace made very little difference to the strength of the associations observed in model 2.

6.2.2 Effect of mode of interview on data quality

The analysis conducted to examine the effect of questionnaire length on data quality was repeated with mode of interview replacing length as the predictor variable. The first stage of our analysis of data quality by mode was to run chi-square tests comparing the marginal distributions of 43 of the variables in the module on well-being (module E).[14] The value for chi-square was statistically significant for 30 out of the 43 items (70%), suggesting that the response distributions differed by mode for most items.[15]

As mentioned, high rates of item non-response can be considered indicative of reduced data quality, although refusing to provide an answer, or 'not knowing' the answer can also both be legitimate responses. In both modes, Don't Know and Refusal options are not explicitly made available to the respondent (that is, they are not read out or presented on the show cards) but they are accepted if the respondent offers them as a response. As with the overall telephone data, item non-response is very low in the face-to-face data, in particular for Refusals and No Answers. The mean rate of Refusals (i.e. the total number of 'Refusals' divided by the number of items in the module) was .13 for the face-to-face group and .00 for the telephone group. The mean rate of No Answers for the face-to-face group was .02 and for the telephone group .00. Finally, the mean rate of Don't Know responses was .66 for the face-to-face group and .26 for the telephone (version A) group. Although use of item non-response is small in both modes, there does appear to be greater incidence in the face-to-face data and this difference is statistically significant for all three types of non-response ($p<0.000$).

Table 30 - Item non-response (mean % of all applicable items in module E)

	Mean		**Sig.**	**N**
	Face-to-face	***Telephone (version A)***		
% of items with refusals	.13	.00	***	5516
% of items with don't know	.66	.26	***	5516
% of items with no answer	.02	.00	***	5516

Notes: ***p<0.001.

As with the earlier analysis, we used the following indicators of satisficing as measures of data quality: *acquiescence, non-differentiation, use of mid-points, use of extreme points*. While the hypothesis for the effect of length of interview on data quality was quite clear, the relationship between mode and propensity to satisfice is more complicated. Previous studies suggest that satisficing is more common in telephone interviews than in face-to-face

[14] TE1-TE3 were not included in the analysis because the variables were incorrectly implemented in the telephone data in some countries. TE47-TE55 were questions asked only of respondents in paid work, which would have led to too small sample sizes in the telephone data.

[15] Because the sample size for the telephone data is not that big (n=**), for some variables the expected cell count was less than 5. Where this affected less than 10% of the cells, nothing was done. Where this affected more than 10% of the cells (te21, te33, te37) categories were combined. Doing this did not change the significance of any of these 3 variables. At TE21 the categories 'most of the time' and 'all or almost all of the time' were collapsed into 'most or all of the time'. At TE33 and TE37 categories 0 and 1, and categories 5 and 6 were combined.

interviews (Holbrook et al., 2003). However, selecting the most 'recently' presented response category is more likely in modes where the categories are offered orally (such as in telephone) while selecting the first offered response option is more likely when the response categories are presented visually (such as in face-to-face). Thus, while the extent of satisficing behaviour is likely to differ by mode, so too is the type. Our null hypothesis is that there will be no relation between mode of data collection and occurrences and type of satisficing.

We first pooled the data in each mode from all countries and ran t-tests to compare the mean scores on the above indicators of satisficing by mode. For items included in each indicator see annex 6. We then repeated the analysis for each country separately.

Table 31 – Mean scores on satisficing indicators – face-to-face vs telephone data

Indicator	Mean		Sig.
	F2F	**TEL**	
Non-differentiation	.53	.54	~
Use of mid-point	.20	.17	***
Acquiescence (agree)	.43	.41	*
Primacy			
Agree strongly	.14	.19	***
Not at all	.08	.08	
Recency			
Disagree strongly	.07	.08	*
A great deal	.11	.17	***

Notes: * p<0.05; ** p<0.01; *** p<0.001; ~ approaching significance

Taking data from all countries together, the summary indicators of satisficing (shown in table 31) were all significantly different by mode except for 'not at all' (*non-differentiation* was approaching significance p=.052). In most cases satisficing was more likely in the telephone questionnaires but use of *mid-points* and '*acquiescence'* were more common in the face-to-face questionnaire. However, pooling the data masked some interesting differences within countries, described here.

Non-differentiation (the tendency to select the same response option on a battery of items) is more likely in the telephone survey when using the summary measure for all countries combined but only for Poland when countries are considered separately. Use of the *mid-point* is more likely in the face-to-face questionnaire. All individual sets of items and the summary measure are significant for all countries combined but when the countries are considered separately, the summary measure is only significant in Switzerland and partly Poland.

Acquiescence (as measured by use of agree) was more likely in the face-to-face questionnaire. However, when looking at the countries separately it is only significant in Poland where it is more likely in the telephone questionnaire.

With regard to use of *extreme points*, 'a great deal' (6 on a 0-6 anchored scale) was more likely in the telephone questionnaire but there was no difference in use of 'not at all' (0 on the 0-6 scale) by mode. Use of strongly agree (the 1st option in a 5-point agree/disagree scale) was significantly more likely in the telephone questionnaire but these relationships were significant only in Switzerland and Germany when countries were considered separately. Use of 'strongly disagree' was more common in the telephone questionnaire, except in Germany where it was more common in the face-to-face questionnaire.

Differences between modes in occurrence of satisficing were most prevalent in Switzerland. In Hungary there was very little difference in the amount of satisficing by mode and the only difference was in use of *extreme points*. In Poland there was no difference between mode for the use of *extreme points*. Only in Poland was the extent of *non-differentiation* significantly different by mode.

As when examining the effect of length on satisficing in the telephone questionnaire, to further analyse the differences between the face-to-face and telephone data we ran regression models, for which the coefficients are shown below in table 32. As well as mode (face-to-face coded 0, telephone coded 1) and country (Germany as reference category), the covariates included were those on which the face-to-face and telephone samples differed significantly (see section 6.1.4), so that we could determine whether differences in data quality were due to differences between the samples or due to mode. These variables were: whether or not the respondent was in paid work (in paid work 1, not 0); area (dummy variables with 'big city' as the reference category); and years of education (high education 1, other 0).

Table 32 – Regression coefficients from OLS model

Indicator	Model 1
Non-differentiation	
- **Constant**	.506***
- **Paid work**	.010***
- **Suburb**	.015***
- **Town**	.007*
- **Village**	.017***
- **Farm**	.010
- **Highed**	-.009***
- **Telephone**	.014**
- **Switzerland**	-.022***
- **Hungary**	.000
- **Poland**	.014***
Acquiescence	
- **Constant**	.401***
- **Paid work**	.025***
- **Suburb**	.022**
- **Town**	.021**
- **Village**	.028***
- **Farm**	.040*
- **Highed**	-.006
- **Telephone**	-.008
- **Switzerland**	-.063***
- **Hungary**	-.119***
- **Poland**	.004
Mid-points	
- **Constant**	.226***
- **Paid work**	-.012**
- **Suburb**	-.010
- **Town**	-.011*
- **Village**	-.013*
- **Farm**	-.005
- **Highed**	-.017***
- **Telephone**	-.018*
- **Switzerland**	-.046***

- **Hungary**	.040***
- **Poland**	.005
Primacy	
Agree strongly	
- **Constant**	.101***
- **Paid work**	.011**
- **Suburb**	.001
- **Town**	.008
- **Village**	.007
- **Farm**	.009
- **Highed**	.012**
- **Telephone**	.031***
- **Switzerland**	.071***
- **Hungary**	.065***
- **Poland**	-.020***
Not at all	
- **Constant**	.093***
- **Paid work**	-.024***
- **Suburb**	-.013*
- **Town**	-.007
- **Village**	-.007
- **Farm**	-.017
- **Highed**	-.021***
- **Telephone**	.001
- **Switzerland**	.028***
- **Hungary**	.099***
- **Poland**	.051***
Recency	
Disagree strongly	
- **Constant**	.088***
- **Paid work**	-.013***
- **Suburb**	-.006
- **Town**	-.005
- **Village**	-.007*
- **Farm**	-.003
- **Highed**	.004
- **Telephone**	.000
- **Switzerland**	.037***
- **Hungary**	.024***
- **Poland**	-.021***
A great deal	
- **Constant**	.078***
- **Paid work**	.002
- **Suburb**	.015
- **Town**	.002
- **Village**	.005
- **Farm**	.000
- **Highed**	.018**
- **Telephone**	.058***
- **Switzerland**	-.003
- **Hungary**	.108***
- **Poland**	.000

Notes: * p<0.05; ** p<0.01, *** p<0.0001

Controlling for differences in the composition of the achieved samples in each mode, we found that mode of data collection had a significant independent effect on the likelihood of satisficing on 4 of the 7 indicators of satisficing. Of those indicators where mode does have a significant effect, telephone respondents were more likely to select the end points 'a great deal' and 'agree strongly' and to not differentiate on scales. Face-to-face respondents were more likely to select the midpoint. So we did find some mode effects but, given the problems inherent in comparing data collected in different modes (see Jäckle, Roberts and Lynn, forthcoming), these findings warrant further investigation.

Being in paid work has a significant effect on propensity to satisfice on all indicators except for 'a great deal'. However, the direction is not consistent. Those in paid work are more likely to use non-differentiation, acquiescence and to agree strongly, but less likely to use mid-points and the scale end-points, not at all and disagree strongly.

Respondents living outside of a big city were less likely to differentiate between scale points on batteries of items and more likely to select the agree option on agree/disagree scales.

High education is significant at all indicators except for disagree strongly and acquiescence. Of those indicators where there is a significant effect, the direction is not consistent. Respondents with high education are more likely to satisfice in terms of using 'agree strongly' and 'a great deal' but less likely to use 'non-differentiation', 'mid-points' and 'not at all'.

7 Additional findings from the CATI study.

In this final section, we summarise the feedback we received from the survey agencies that participated in the CATI study about their experiences of carrying out the ESS by telephone.

The overall impression from the participating survey agencies, with the exception of Cyprus, was that CATI data collection would be possible on the ESS, but that the response rate would be lower than that for face-to-face. Although interviewers found the survey quite difficult to 'sell', most found the interviews (all versions) fairly easy to administer and both the respondents and the interviewers found it interesting. In other words, although it was difficult to persuade the respondents to take part, once they did, the interview went smoothly. This was mainly put down to the interesting and varied topics.

Call record forms
The survey agencies reported that the call record forms were more complicated and had more codes than they and the interviewers were used to. As a result additional training of interviewers was needed. Some specific problems with the call record forms were also noted; for example that some interviewers had difficulty matching some calls to the codes, and some felt the list of codes could be simplified.

Participation
In the project instructions we included an example of a possible introductory text that could be used by interviewers, as well as suggesting the key information that should be included in the introduction. It seems that survey agencies may have misunderstood that the text was simply an example and adopted it word-for-word. A number of survey agencies reported that the text was too long and didn't 'sell' the questionnaire well, even going so far as to suggest that it may have had a detrimental effect on the response rates.

Overall, interviewers found version A and part 2 of version C the most difficult to 'sell' and part 1 of version C the easiest. Respondents didn't react well to part 2 of version C and interviewers felt uncomfortable about 'lying' to the respondent (i.e. by not disclosing that there would be a second interview at the start of the first). The length of the interview was the reason most often cited by interviewers for 'difficulties when administering the interviews' for version A, while 'encouraging the respondent to participate in the second part' was the reason most often cited for version C.

The Polish survey agency recommended not leaving a voicemail if no-one answered. It was thought that it might increase refusals since respondents might make up their mind about whether or not to participate and by the time the interviewer contacts them it is too late to convert them. However, offering to conduct the interview at another time was helpful, as was offering to complete it at another time. Making appointments was found to be useful by some, although others suggested sample members were using 'making an appointment' as a way of refusing. Appointments also require central organisation, for example to make sure an interviewer is available and calls back at the correct time.

Data quality
Signs of fatigue were observed by the interviewer least in part 1 of version C and most in version A (after approximately 30 minutes), and were observed most of all among elderly respondents. Signs of 'irritation' were less common than signs of fatigue but were observed most in version A (after approximately 45-50 minutes). Most survey agencies reported particular problems with module D (the questions are very repetitive and differentiating between 'your' opinion and 'most people's' opinion was problematic for many) and module E (repetition of the same scales). 11-point scales and long lists of responses were hard for respondents and the long lists often had to be repeated. The questions about parents' occupation and education were found to be difficult for older respondents.

A problem was identified with the wording of TF78 (*Now we have finished the interview, I just want to ask you about the length of the interview. Would you have been willing to continue ... much longer, a bit longer, or not at all?*) as some respondents thought if they said 'much longer' or 'a bit longer' that the interview would continue.

Swiss incentives experiment and refusal conversion
The Swiss survey agency conducted an incentive experiment offering respondents either 20CHF or 30CHF. Higher response rates were achieved in the 30CHF group and the effect of increasing the incentive was bigger the shorter the questionnaire. So the larger incentive actually increased the difference in response rates between the 3 versions of the questionnaire.

Refusal conversion efforts were found to be more successful the longer the questionnaire. The conversion rate was higher in the higher incentive group.

8 Discussion

In this paper we report on an experiment carried out in the context of the European Social Survey, designed to examine the effect of inviting respondents to participate in telephone interviews of different lengths on their willingness to participate in the survey. Three treatment groups were interviewed with three versions of the ESS questionnaire, adapted for telephone administration. One version was identical in format and length to the standard face-to-face interview and lasted around 1 hour (version A), one contained just one of the two rotating modules reducing the length to 45 minutes (version B) and the other was the full ESS questionnaire divided in half to be administered in two parts of 30 minutes each (version C).

Respondents to version C were asked to participate in a 30 minute interview, at the end of which they were asked to participate in *another* 30 minute interview – either straight away or at another time. As a result of the changes that had to be made to the questionnaires to alter the length, one module (module E) appeared in a different position (that is it was preceded by a different number of items) in each version. This enabled us to also consider the effect of length on the quality of data collected by comparing the responses in module E in each group. Although not the main aim of this experiment, we also conducted some comparisons between the telephone data and data from the round 3 ESS face-to-face survey in order to identify differences in response rate and data quality by mode of data collection.

8.1 Summary of findings and discussion

8.1.1 Rates of participation

For the analysis of rates of cooperation we used data from the call records and presented comparisons on overall and full rates of cooperation, refusal rates and overall and full response rates, to assess the impact of interview length on survey outcome. Our hypothesis was that the response rates would be lower for the longer questionnaires. We also examined the extent to which the achieved samples in each group varied on a range of socio-demographic variables, to evaluate whether interview length would have a differential effect on the willingness to respond of different subgroups of respondent. Finally, we compared the response rates and sample composition from the hour-long telephone questionnaire with those of the round 3 face-to-face survey in order to consider possible differences that might occur in propensity to respond and resulting non-response bias due to mode of data collection. Particularly due to the length of the questionnaire we would expect a higher response rate in the face-to-face survey.

The results of the experiment broadly confirmed the prediction that the announced length of the survey interview can influence the target respondent's willingness to participate. Differences in levels of cooperation were observed between the treatment groups, although the pattern of findings was not always consistent. In almost all cases, the '30 minute' interview attracted higher proportions of respondents than the 60 minute interview. However, the 15-minute difference between groups A and B and groups B and C did not always lead to differences in rates of cooperation, and where there were differences they tended to be small.

Overall, respondents in group C were more likely to agree to participate, but because their full cooperation was dependent on them agreeing to take part in *two* 30 minute interviews, the apparent advantage of the version C interview in terms of overall response rates was not also reflected in the full response rate. After taking into account partial interviews, break-offs and those unwilling to complete part 2, response rates were lowest overall in group C compared to the other groups. Feedback from the survey agencies supported this finding as version C was the least popular among interviewers. Based on these findings, we would conclude that in terms of securing the cooperation of respondents, interview length is an important predictor in determining willingness to participate in a survey, and that the shorter the interview the better (of the lengths that we tested). If the aim, however, is to administer a long survey questionnaire (as is the case on the ESS), then there appears to be no clear advantages to splitting the questionnaire into two (or more) parts, at least not in the way that we did it here. That is to say, these somewhat negative findings are likely to have resulted from the way in which group C respondents were invited to take part in the second part of the interview (at the end of part 1) having initially been told that it was a 30 minutes questionnaire (see section 6.2.4). There may, however, be considerable advantages to offering respondents the chance to complete an hour-long interview in multiple parts (at their discretion), having first secured their cooperation.

The number of break-offs and partial interviews (that is, the difference between the full and overall response rates) was quite small overall, suggesting that once interviewers persuaded the respondent to take part, they were willing to continue until the end. Feedback from the survey agencies supported this finding, saying that, although getting the respondents to take part was difficult, once they did agree to do it, it was relatively easy to keep them engaged and the interviews themselves were straightforward and unproblematic. This is likely to be to do with the range and subject of the topics covered in the European Social Survey which, being varied, seem to help keep the respondent interested and engaged.

Some interesting differences in the results of the experiment were observed across the participating countries. These differences may well be informative of cross-cultural variations in the acceptability of long telephone interviews and, therefore, the suitability of telephone interviewing as a mode of data collection in those countries. This difference should not be surprising. We know, for example, that some survey agencies set a limit on the length of interview they will conduct by telephone, usually ranging between 20 and 30 minutes, while others don't (see Roberts, Eva and Widdop, 2008). However, the results from Cyprus give us reason to be cautious in our interpretation of the differences between countries. In particular, despite our best efforts to control the way in which the experiment was implemented in each country (e.g. by strictly specifying the protocol for introducing the survey and the call recording procedures), there may still have been differences in the implementation of the study that could account for the somewhat varied findings. Feedback from interviewers and survey organisations participating in the study confirmed that neither had much prior experience of conducting a telephone survey with such long questionnaires. Similarly, almost all experienced difficulties with maintaining the call records as specified by the project team (notably, the list of outcome codes provided were not always compatible with how the existing CATI program had been programmed and was found to be overly long for the interviewers coding the call outcomes - see section 7). As a result, there were some differences in how the outcome codes were used by the interviewers and this may be reflected in the reported outcomes.

In all countries the response rate achieved in the round 3 face-to-face survey was higher than the hour-long telephone survey. These differences were large in Poland, Hungary and Germany. The source of these differences was not the same in each country. Hungary and Germany had much higher refusal rates in the telephone survey than in the face-to-face, whereas in Poland the difference in refusal rate was quite low between modes. In Switzerland the full round 3 face-to-face response rate was only 8 percentage points higher than the telephone response rate.

Almost no differences were found between the sample composition in each of the 3 treatment groups, with no significant differences in Germany and Switzerland. Where differences were found they were in group C: in Hungary there were significantly more respondents in paid work in group C and in Poland there were significantly fewer male respondents.

There was mostly no difference between the face-to-face and telephone samples, although years in education and paid work differed significantly in almost all countries. Years of education was significantly higher in telephone in all countries except Switzerland. The proportion in paid work was higher in the face-to-face sample in all countries except Hungary. These results show some support for the hypothesis that the people who respond to telephone surveys differ in systematic ways from those who respond to face-to-face surveys. In what ways and to what extent they differ varies across countries, although the effect of 'high education' and being in paid work in the face-to-face data is fairly consistent.

8.1.2 Data quality

In order to measure the data quality we looked at the similarity of marginal distributions at the item level, the rate of missing data in the module and the extent of respondent satisficing. In each case, our prediction was that data quality would be poorer in the longer questionnaire condition due to the increased cognitive burden on respondents of answering lots of questions and declining motivation over the course of the interview. The average length of interview across treatment groups was as we expected (1 hour, 45 minutes, 2*30 minutes).

As predicted, the analysis of data revealed significant differences in the quality of the data across the three treatment groups. The results of the chi-square tests found around one quarter of the items revealed significant differences in response distributions and in almost all cases, it was responses in group C that differed most from those given by groups A and B. Item non-response rates across the three groups were very low and broadly similar (though there were slightly fewer Don't Know responses in group C compared with groups A and B). To test our hypothesis in relation to satisficing, we used both ANOVA test and OLS regression models to assess the effect of varying the length of the questionnaire on the tendency to adopt particular response sets: non-differentiation, acquiescence, preference for middle alternatives and response order effects in rating scales. The results of our analysis were mixed.

There were no differences between the three groups in the overall extent of non-differentiation. However, we did find differences in the tendency to select specific scale points to respond to a battery of items all rated on the same scale. Notably, we found greater preference for the use of end-points (both primacy and recency effects) among respondents in the long questionnaire group compared with those in the short questionnaire group. The direction of the effect differed depending on the type of scale (in group A, primacy effects were more common on the fully labelled 'agree strongly to disagree strongly' scale, while recency effects were more common on the 7-point anchored scale). By contrast, respondents in the shorter questionnaire group showed more preference for mid-points and were more likely to select the 'agree' response, which could be indicative of acquiescent response style. Though mixed, the effects were quite robust, and even when controlling for sex, age, and country, the pattern of results relating to the questionnaire length treatment was largely unchanged. The exception was acquiescence, for which the effect of being in the short questionnaire group dropped out once controls for country were added to the model. Controlling for the pace of the respondent's interview did not affect the results of the models.

In summary, the results suggest that the effect of questionnaire length on response distributions may have been restricted to only two types of rating scale: 5-point agree/disagree scales and 7-point semantic differential scales, with respondents answering the long questionnaire showing a preference for end-points and respondents answering the short questionnaire showing a preference for mid-points. This mixed pattern of effects provides some limited support for our hypothesis that answering long telephone interviews encourages respondents to satisfice. This conclusion is further supported by feedback from the survey agencies suggesting that signs of fatigue were most common in version A.

In terms of the mode comparisons, the overall interview length was found to be roughly the same in the telephone and face-to-face interviews. The average was slightly longer in face-to-face than telephone in all countries except Poland. The majority of items were found to significantly differ by mode (when differences between samples were not controlled for) and although item nonresponse was very low in both modes it was significantly more common in the face-to-face survey. However, although very high levels of item non-response would suggest poor data quality, it is not necessarily the case that less item non-response is always better than more since 'don't know' and 'refusal' can be valid responses and may be

preferable to respondents giving a 'false' answer. When controlling for differences between the samples, satisficing overall was more likely in the telephone questionnaire, but preference for mid-points was more likely in the face-to-face interviews. Feedback from the interviewers suggested that questions with long response scales were problematic in the telephone survey with no showcards, which may help to explain the increased tendency for telephone respondents to satisfice.

While the headline findings support our expectations that satisficing would be more likely in long telephone interviews, for both sets of comparisons, the inconsistency of the findings makes it difficult to draw conclusions about the mechanisms by which interview length and mode exert an influence on response quality. Part of the difficulty may stem from a lack of clarity over what constitutes better quality data and this again may be partly due to the somewhat mixed evidence that the response effects investigated here can be attributed to respondent satisficing. In particular, studies looking at predictors of selecting middle alternatives have not always produced findings that are consistent with the theory, which may suggest that the greater use of this response option among group C telephone respondents ought not to be attributed to shortcutting (it may indeed be indicative of increased validity; or alternatively, suggest a more cautious response style). Interestingly, use of mid-points behaved differently from the other indicators of satisficing in our comparisons of mode and interview length. Only use of mid-points was more common in the short questionnaire and the most common in face-to-face.

8.2 Problems with the design

There were a number of limitations of the design of our study that both restricted the analysis that we were able to do and should give us reason to be cautious about how we interpret some of our findings.

8.2.1 Overall

The context to this study is a programme of research investigating the possibility of alternative modes of data collection on the ESS. Previous research found that telephone interviewing is the mode most likely to be used in an alternative data collection design (Roberts, Eva and Widdop, 2008). This experiment aimed to investigate the feasibility of conducting the ESS by telephone. Although our choice of countries to include in this experiment was restricted by available budget and by fieldwork agencies' capabilities and availability at the time of the round 3 fieldwork, we did aim to represent the range of countries within the ESS, in terms of differing survey traditions and data collection challenges. However, another approach would have been to conduct the experiments in the countries most likely to want to, and/or be able to, adopt telephone interviewing on the ESS, such as Sweden. Unfortunately the countries that are interested in switching to telephone interviewing tend also to be those where survey fieldwork is most expensive and this approach would have reduced the number of countries that we could have included in our study.

While the selection of countries may affect the usefulness of the findings for the ESS project, a problem with the experimental design in terms of applicability to the wider research world is that all three questionnaires that we tested would be considered long for telephone administration. Short telephone surveys are much more common (in fact many survey agencies limit telephone interviews to under 30 minutes (Roberts, Eva and Widdop, 2008)) and it is possible that bigger differences in response rates might be observed if we tested a questionnaire under 30 minutes. It is unlikely, therefore, that the results of our experiment are

transferable to shorter questionnaires, though our research provides quite a robust test of the challenges involved in conducting long survey interviews such as the ESS by telephone.

8.2.2 Rates of participation

The main problem we encountered when testing the effect of length on rates of participation was the operationalisation of the call record forms, which were longer and more complex than most survey agencies and interviewers were used to (see section 7). This may have affected the equivalence of the call record data from the five countries. However, this is in itself an interesting finding as the possible problems of implementing the call record forms has implications for a move towards telephone data collection on a survey like the ESS where the need to collect detailed paradata for the purpose of calculating response rates and assessing nonresponse bias is essential.

The main problem when comparing the effect of mode on rates of participation was that the same amount of effort was not put into the telephone survey as the face-to-face survey. For example, only one country used advance letters and incentives in the telephone survey, whereas almost all did in the face-to-face survey (see section 4.3.2). It is known that these kind of response enhancement techniques can have an effect (sometimes an important effect) on response rates so we must consider the fact that data collection efforts were not comparable by mode in all cases which may go some way towards explaining why the response rates achieved in the telephone survey were so much lower than in the face-to-face survey. It is likely that in some countries, for example Switzerland, if they conducted the actual ESS by telephone they could achieve a higher response rate than they did in our study, by spending more, while still spending far less than they would on the face-to-face survey (although Switzerland is the one country that did use both an advance letter and incentives).

8.2.3 Data quality

The measurement of data quality was not one of the initial aims of this study, which was designed to assess the effect of informing target respondents of the likely length of the questionnaire on response rates, and this had a number of ramifications for the data quality analysis that we conducted. Firstly, in the telephone study, the respondents in each group cannot strictly be considered to have been randomly assigned to the treatments. Respondents effectively selected themselves into the groups by agreeing to participate meaning that propensity to satisfice was unlikely to have been evenly distributed between the groups to begin with (to the extent that it is related to overall motivation to respond)[16]. Similarly, there is a selection effect in the mode comparisons since we might expect different people to respond to different modes (though we explicitly controlled for this in our multivariate analyses). Response rates were lowest in group A and for both parts of group C, suggesting that sample members in these groups were the most reluctant to participate. Respondents in group C were not forewarned of the second part interview, so those that did respond to both parts may have been especially reluctant to respond diligently by the time they were responding to part 2. Alternatively, those who responded to both parts of version C might be particularly motivated respondents. It is not clear to what extent and how this weakness of the study may have impacted on the results, but it may provide some explanation for the mixed pattern of observed effects.

Secondly, the number of questions preceding the module of questions analysed was manipulated by changing the overall order of the questionnaire (as well as cutting a module altogether from version B). This meant that the context in which the questions on wellbeing were presented was different for group A (where the module was preceded by a module of

[16] We have investigated and attempted to address this problem elsewhere (see Roberts, Eva, Allum and Lynn, 2008).

questions on the timing of life events) to that in groups B and C (where the module was preceded by core questions on well-being and ethnic and national identity). This alone could account for the differences observed between groups A and C, although the presence of differences between groups B and C should give some reassurance that our initial conclusions are justified. In addition, there may be an effect of the version C respondents having previously answered another 30 minute survey, albeit on a different occasion. This was not an issue for the telephone/face-to-face comparisons since the question order was the same. Though we have not done so here, one possible way of testing this would be to see if there were differences in satisficing across items appearing earlier in the questionnaire.

There are other reasons why the module of questions we focused on here (module E) may not have been well-suited to testing the data quality hypothesis we were interested in, in the way we wanted to test it. In particular, it is reasonable to assume that of all the topics in the ESS, respondents may have most enjoyed answering the questions in module E which asked respondents about themselves and how they were feeling. This may have encouraged more careful responding than modules covering less involving topics. However, feedback from the survey agencies suggested – as we anticipated - that respondents found the repetition of the same scales in module E boring. Furthermore, the repetitive nature of the scales in this module to a certain extent limited the scope of our analysis, restricting us to just a few response scales on which to examine a range of different effects.

8.3 Recommendations for the ESS

This experiment showed that in some aspects, telephone as a mode of data collection could be feasible on the ESS. There were low levels of break-offs, low item non-response, and interviewer feedback suggested that, although persuading respondents to take part was harder than in face-to-face, conducting the interview by telephone was unproblematic. It should also be noted that there was some success of the telephone experiment in terms of response rates achieved versus face-to-face, although this differed across countries. In Round 1 the response rate in Switzerland was actually lower than the full response rate of the telephone version A. It has only increased in later rounds due to extensive efforts and incentives, which are expensive.

However, in other areas, telephone as a mode of contact and data collection remains problematic. No country has complete coverage (and coverage of telephones that can be sampled is falling in many countries) and lower response rates in most countries. The collection of call record data also proved problematic for all countries (see section 6.2.4) since the call records were longer and more complex than they were used to. For most surveys, including the ESS, the collection of this type of paradata is an important part of the methodology that enables full analysis of response rates, response sequences, the effect of response enhancement techniques, and so on. If it proved impossible to collect this type of data in a standardised way across all countries for telephone data collection, this would seriously influence whether or not data collection by telephone could be considered on the ESS. Furthermore, even if standardised paradata could be collected, some reduction in the quality of call record data for data collection by telephone is unavoidable since it is not always possible to know what the correct outcome code is, for example whether a case is ineligible or non-contact. It is these areas that seem to suggest that although telephone would be possible as a additional mode of data collection, its suitability as a uni-mode design on the ESS is limited.

As well as testing the feasibility of conducting the hour-long ESS by telephone, this study also tested possibly alternatives if an hour proved impossible. Although the response rates for

version B (45 minutes) were better than version A (1 hour), the differences were minor (1.6 to 4.9 percentage points). Sizeable improvements in response rates did occur for the 30-minute questionnaire but the response rates for the full questionnaire split in 2 was the lowest of all. So in practical terms, there is no strong argument for either splitting the ESS questionnaire (at least not in the way we did in this experiment) or reducing it by 15 minutes. (Although there may be advantages to informing the respondent that they can complete the interview over more than one occasion if they would prefer – a finding that was backed up by feedback from the survey agencies.) In addition, version C was also the most different in terms of data quality and again the differences between versions A and B were minor. And so, in terms of data quality there is also not much to be gained from reducing the questionnaire by 15 minutes. Furthermore, since the suggested design for the 45-minute version involved asking each rotating module of only half the sample, the sample size would have to be increased. It is likely that the resulting increased costs could not be justified by the minor improvement in the response rates achievable by telephone.

With regard to the effect of mode on data quality, this study, like others, found that there were some differences between telephone and face-to-face interviews, which suggests that comparing data from different modes cannot be done without caution. However, currently on the ESS there are various issues that may affect the comparability of the data between countries, in particular the very varying response rate. Currently analysts are comparing data from countries with response rates that vary by as much as 27% (round 3), 35.5% (round 2) and 46.5% (round 1).

On the basis of the findings from this study, we make the following tentative recommendations regarding future data collection strategies on the ESS. First and foremost, if any change away from face-to-face as the sole mode of data collection is to be made, a mixed mode design would seem to be the most promising alternative. The design should use single modes of data collection for each respondent but more than one mode within countries and the design used may differ in each country or may be standardised across all countries.

Introducing a mixed mode data collection design might improve response rates in those countries that are particularly struggling and as a result might make the response rates more comparable across countries. Furthermore, and more importantly, mixed mode designs might achieve a more representative sample in those countries where certain sub-groups are currently under-represented. This is particularly possible if introducing a cheaper mode of data collection allows countries to spend more on response enhancement techniques. So although mixed mode data collection would certainly have negative impacts in some respects, these may be off-set by advantages in others.

Overall, it would seem that telephone is not well-suited as a single mode option for a survey like the ESS, but it may be of value as a supplement to other modes within countries. It is important to note that results differed across countries. As a result, not only should decisions about whether or not to mix modes be taken on a country-by-country basis, each country should also be required to conduct their own experiment to test the feasibility of a mixed mode design in their country before a decision is made on whether or not a mixed-mode design could be implemented. In terms of both response rates and data quality, there is not much to be gained from reducing the length of the interview, and so if telephone were added as a mode of data collection on the ESS, it seems that the full hour-long questionnaire could be used.

8.4 Avenues for future research

The future work that could be carried out following on from this experiment can be split into two parts: additional analysis of the telephone experiment data; and future topics for research.

8.4.1 Additional analysis

One possibility for further analysis of these data would be to explore in more detail the extent to which response effects vary in each of the participating countries. Based on our knowledge of cross-national variation in survey practice (see Roberts, Eva and Widdop, 2008), we had expected that interview length would exert an influence on response propensities differentially across countries, due to national variation in tolerance for long interviews. This was not borne out in the data. That is to say, although the response rates achieved differed across countries, the overall effect of length on response rates generally did not. However, we did see differences by country in the direction and extent of different response effects independent of the effects of questionnaire length that we have not attempted to unravel here. A deeper exploration of the different predictors of satisficing in each country may shed some light on the nature and causes of the effects observed here.

Another future area of analysis could involve a more detailed processing of the call record data, to assign final outcome codes to sample cases, based on hierarchical ordering of the outcome codes allocated at different contact attempts. The present analysis was based on the most recent case dispositions (i.e. the outcome of the last call attempt), and in this respect, our results are unlikely to provide the most accurate representation of the survey outcomes. Though previous studies (e.g. McCarty, 2003) suggest the differences in response rates reported as a function of different methods of assigning final outcome codes are unlikely to be large, it is clear that precision can be gained by a more thorough analysis of the data available to us.

We found that the measures of data quality we used behaved in different ways and further analysis could be done to investigate the cause of this. For example, our analysis suggests that the likelihood of respondents selecting either the first or last response option may differ depending on the type of scale. In the longest questionnaire respondents were more likely to select the first option of the fully labelled agree strongly to disagree strongly scale, but the last option on the 7-point anchored scale. In addition, we were not able to test primacy/recency effects on a list of categorical response options, which may provide different results again.

8.4.2 Future topics for research

There is still a lot of research that needs to be done and questions that need to be answered before the ESS could consider a move to alternative modes of data collection, including: investigating the effect of the released data having been collected in multiple modes for the data user; gathering more information on how respondents and interviewers use showcards and the effect they have on response quality; examining the occurrence of social desirability in different modes as an alternative measure of data quality; looking at ways to correct for mode effects that cannot be mitigated through the design of more equivalent questionnaires. Two issues that have arisen from this study are discussed in more detail here.

Firstly, one of the key issues that still needs to be answered is how to design data collection instruments in different modes to mitigate mode effects. We would need to investigate the effect of changing the questions on the response distributions, especially for those items that had to be changed substantially. Although no experimental design was included in this study to compare different question versions, we could in part do this by comparing the round 3 face-to-face data with the telephone data. We would need to control for sample composition to allow for the fact that different types of people might respond to different modes (see

section 6.1.4 for comparisons of sample compositions by mode). However, one problem with this is that the variables we would need to control for are also the variables that had to be changed the most (e.g. questions about income, education, employment), meaning there may be mode effects (or question form effects) affecting the comparability of our control variables. On the other hand, since the questions that need most altering (socio-demographic variables with complex response options) tend to be factual or behavioural questions, they might also be the ones that are least susceptible to mode effects.

Although the version C questionnaire was not successful in terms of full response rates, we did find that a number of respondents completed the questionnaire over multiple occasions[17]. If this is an inevitable consequence of conducting a long questionnaire by telephone we would need to find out more about its effects on data quality. It would also be interesting to investigate, for matters of equivalence, whether or not respondents to the face-to-face survey complete the survey in multiple parts. This is likely to be less common than on telephone interviews which respondents find easier to stop (at which point the interviewer is likely to offer them the chance to complete the survey at another time) as well as being harder to organise from a logistical point of view and more expensive.

Ultimately the next stage of the mixed mode programme of research would be to test a full mixed mode design. There are a number of findings from this study, as well as from other research carried out within JRA1, that would lead us to consider this. This experiment revealed that conducting the ESS by telephone is feasible but, due to insufficient coverage and poor response rates, it could not be used as the sole mode of data collection for the ESS. This supports our findings from the Mapping Exercise (Roberts, Eva and Widdop, 2008) that found that face-to-face is the only workable single mode design possible for the ESS, perhaps unsurprisingly, given that the survey was originally conceived as and designed for face-to-face administration. As a result any change in mode of data collection for the ESS seems likely to be towards a mixed mode design, which would be likely to entail some redesign of the original face-to-face instrument. For this reason, the next stage of research should answer questions such as, what type of mixed mode design would be most appropriate, which modes would be included, and what would be the implications for cost, response rates and data equivalence, as well as address the issue of how to adapt the questionnaires to maximise comparability across alternative modes.

[17] These individuals could not be included in the analysis of data quality

9 References

The American Association for Public Opinion Research (2008). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*, 5th edition. Lenaxa, Kansas: AAPOR

Bradburn, N.M (1978) 'Respondent Burden'. *ASA proceedings of the Survey Research Methods Section,* 35-40

Clancy, K.J. and Wachsler, R.A. (1971) Positional effects in shared-cost surveys. *Public Opinion Quarterly*, 35, 258-265.

Clarke III, I. (2001) Extreme response style in cross-cultural research. *International Marketing Review*, vol. 18, no. 3, 301-324.

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.

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

Frankel, J & Sharp L.M (1981) 'Measurement of Respondent Burden.' *Statistical Reporter* 81(4) 105-111

Collins M, Sykes W, Wilson P and Blackshaw N. (1988). 'Nonresponse: The UK Experience.' *In*: *Telephone Survey Methodology*, Groves et al (eds). Wiley: New York. 213-232

Groves, R and Lyberg L (1988) 'An overview of nonresponse issues in telephone surveys.' *In*: *Telephone Survey Methodology*, Groves et al (eds). Wiley: New York. 191-212

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

Hansen, K (2006) 'The Effects of Incentives, Interview Length, and Interviewer

Characteristics on Response Rates in a CATI-Study'. *International Journal of Public Opinion Research* 19(1) 112-121

Heberlein, T and Baumgartner, R (1978) 'Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature.' *American Sociological Review*. 43(4) 447-462

Herzog, A.R. and Bachman, J.G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, 45, 549-559

Holbrook, A. L., Cho, Y. I., & Johnson, T. P. (2006). *Extreme response style: Style or substance.* Paper presented at the Annual meeting of the American Association for Public Opinion Research, Montreal, Canada.

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.

Hui, H.C. and Triandis, H.C. (1989) Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, vol. 20, no. 3, 296-309

Jabine, T., Straf, M., Tanur, J. and Tourangeau, R. (1984). *Cognitive aspects of survey methodology: building a bridge between disciplines.* Washington, DC: National Academy Press.

Jäckle, A., Roberts, C. and Lynn, P. (Forthcoming) Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review.*

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.

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

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

Kraut, A.I., Wolfson, A.D. and Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, 60, 774-776.

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., & 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. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567. Krosnick, J.A., Judd, C.M., and Wittenbrink, B. (2005). The measurement of attitudes. In

D. Albarracín, B.T. Johnson and M.P. Zanna (Eds.). *The handbook of attitudes*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in Survey Research* (Vol. 70, pp. 29-44). San Francisco: Jossey-Bass Publishers.

Lynn, P., Beerten, R., Laiho, J. and Martin, J. (2001) Recommended standard final outcome categories and standard definitions of response rate for social surveys. *ISER Working Paper,* 2001-23, Colchester: University of Essex

Marquis, K 1979 'Survey Response Rates: Some Trends, Causes and Correlates.' *Health Survey Research Methods*. Seeona Biennial Conference, DHEW Publican No. (PHS) 79-3207, National Center for Health Services Resarch, Hyattsville, MD.

McCarty, C. (2003). Differences in response rates using most recent versus final dispositions in telephone surveys. *Public Opinion Quarterly*, Vol. 67; 396-406.

Meegama, N and Blair, J, 1999(?) The Effects of Telephone Introductions on Cooperation: An Experimental Comparison. Unpublished?

Morton-Williams, J and Young, P (1987) 'Obtaining the Survey Interview – and Analysis of Tape Recorded Doorstep Introduction.' *Journal of Market Resrach Society*. 29(1) 35-54.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, Vol. 60*, 58-88. Roberts, Eva and Widdop, 2008

Philippens, M., Billiet, J., Loosveldt, G., Stoop, I. and Achim, K. (2003) *Non-response and fieldwork efforts in the ESS: Results from the analysis of call record data.* Work Package 7 – Data-based quality assessment in the ESS Round 1 (part II).

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., Eva, G., and Widdop, S. 2008. *Assessing the Demand and Capacity for Mixing Modes of Data Collection on the European Social Survey: Final Report of the Mapping Exercise*. Unpublished manuscript, available on request from the Centre for Comparative Social Surveys, City University, London.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic. Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

10 Annexes

Annex 1. Question wording and response options for module E.

No.	Question Wording	Response Categories	X^2	d.f.	p
TE1	In the past 12 months, how often did you get involved in work for voluntary or charitable organisations?	1=At least once a week 2=At least once a month 3=At least once every three months 4=Less often 5=Never	7.27	8	.51
TE2	How often, in the past 12 months, did you actively provide help for other people?	1=At least once a week 2=At least once a month 3=At least once every three months 4=Less often 5=Never	17.31	8	.03
TE3	And in the past 12 months, how often did you help with or attend activities organised in your local area?	1=At least once a week 2=At least once a month 3=At least once every three months 4=Less often 5=Never	6.11	8	.64
TE4	"I'm always optimistic about my future."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	4.85	8	.77
TE5	"In general I feel very positive about myself."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	10.12	8	.26
TE6	"At times I feel as if I am a failure."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree	13.97	8	.08

		5=Disagree strongly			
TE7	"On the whole my life is close to how I would like it to be."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	10.47	8	.23
TE8	How much of the time during the past week did you feel depressed?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	6.55	6	.37
TE9	How much of the time did you feel that everything you did was an effort?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	3.10	6	.80
TE10	How much of the time during the past week was your sleep restless?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	5.02	6	.54
TE11	How much of the time did you feel happy?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	8.93	6	.18
TE12	How much of the time during the past week did you feel lonely?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	3.24	6	.78
TE13	How much of the time did you enjoy life?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	6.56	6	.37
TE14	How much of the time did you feel sad?	1=None or almost none of the time 2=Some of the time 3=Most of the time	8.53	6	.20

		4=All or almost all of the time?			
TE15	How much of the time did you feel you could not get going?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	18.35	6	.005
TE16	How much of the time did you have a lot of energy?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	6.13	8	.63
TE17	How much of the time did you feel anxious?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	5.78	6	.45
TE18	How much of the time did you feel tired?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	9.27	6	.16
TE19	How much of the time were you absorbed in what you were doing?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	6.32	6	.39
TE20	How much of the time did you feel calm and peaceful?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	4.00	6	.68
TE21	How much of the time in the past week did you feel bored?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	4.47	6	.61
TE22	How much of the time did you feel rested when you woke up in the morning?	1=None or almost none of the time 2=Some of the time 3=Most of the time 4=All or almost all of the time?	5.34	6	.50

TE23	"I feel I am free to decide for myself how to live my life."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	13.23	8	.10
TE24	"In my daily life, I seldom have time to do the things I really enjoy."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	10.52	8	.23
TE25	"In my daily life I get very little chance to show how capable I am."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	15.27	8	.05
TE26	"I love learning new things."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	16.49	8	.04
TE27	"Most days I feel a sense of accomplishment from what I do."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	8.00	8	.43
TE28	"I like planning and preparing for the future."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	17.37	8	.03
TE29	"When things go wrong in my life, it generally takes me a long time to get back to normal."	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree	12.47	8	.13

		5=Disagree strongly			
TE30	“My life involves a lot of physical activity.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	7.61	8	.47
TE31	How satisfied are you with how your life has turned out so far?	0=Extremely dissatisfied 10=Extremely satisfied	23.10	18	.19
TE32	How satisfied are you with your present standard of living?	0=Extremely dissatisfied 10=Extremely satisfied	45.33	20	.001
TE33	How much of the time spent with your immediate family is enjoyable?	0=None of the time 6=All of the time	12.50	12	.41
TE34	How much of the time spent with your immediate family is stressful?	0=None of the time 6=All of the time	22.75	12	.03
TE35	To what extent do you get a chance to learn new things?	0=Not at all 6=A great deal	19.86	12	.07
TE36	To what extent do you feel that people in your local area help one another?	0=Not at all 6=A great deal	18.40	12	.10
TE37	To what extent do you feel that people treat you with respect?	0=Not at all 6=A great deal	12.21	12	.43
TE38	To what extent do you feel that people treat you unfairly?	0=Not at all 6=A great deal	27.32	12	.007
TE39	To what extent do you feel that you get the recognition you deserve for what you do?	0=Not at all 6=A great deal	13.74	12	.32
TE40	“I generally feel that what I do in my life is valuable and worthwhile.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	16.03	6	.01

TE41	“If I help someone I expect some help in return.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	18.18	8	.02
TE42	“The way things are now, I find it hard to be hopeful about the future of the world.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	8.35	8	.40
TE43	“There are people in my life who really care about me.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	8.77	8	.36
TE44	“For most people in [country] life is getting worse rather than better.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	8.21	8	.41
TE45	“I feel close to the people in my local area.”	1=Agree strongly, 2=Agree, 3=Neither agree nor disagree 4=Disagree 5=Disagree strongly	13.50	8	.10
TE46	How often, if ever, do you feel frustrated by having watched too much television?	1= Often 2=Sometimes 3= Occasionally 4= Never 5=Or, do you never watch TV?	17.12	8	.03

Annex 2. Outcome codes

A. Call outcome codes for samples based on lists of households/numbers (Cyprus and Switzerland)

1. Contact, interview:

1.1. Complete Interview (for group C, this means both interviews complete)
1.2. Partial interview, i.e. break-off (for group C, this means complete part A and partial part B)

GROUP C ONLY:
1.3. Complete interview part A
1.4. Partial interview/break-off part A

For any call resulting in partial interview, interviewers should also record whether it is recommended to call again to complete the interview (e.g. circumstantial inability to continue the interview) or not (e.g. hard refusal). If refusal, also answer questions 1 to 4 below.

2. Non-contact:

2.1. Ring, no answer after 7 rings
2.2. Busy/engaged line
2.3. Answering machine/voice mail; message left
2.4. Answering machine/voice mail; no message left
2.5. Disconnected or other non-working
2.6. Fax/modem/data line/pager
2.7. Telecommunication technological barriers e.g. call barring, call screening
2.8. Technical phone problems e.g. bad line, Telephone Company experience technical problems
2.9. Call forwarded, no answer
2.10. Phone number has changed – new number ascertained
2.11. Phone number has changed – new number not ascertained
2.12. Incorrect number (e.g. number doesn't match address)
2.13. Other non-contact - Give full details.

3. Contact – no interview:

3.1. Contact made at given phone number but selection procedure not completed; phone back at another time (not including if it is found that the household has moved: see 3.11 and 3.12)
3.2. Contact made, selection procedure completed but no contact with sample person
3.3. Contact made with sample person; appointment made
3.4. Contact made with sample person; phone back at another time
3.5. Contact made with sample person; other. Give details.
3.6. Sample person temporarily unavailable
3.7. Language barrier
3.8. Sample person unavailable (e.g. away, in hospital)
3.9. Sample person ill at home
3.10. Sample person unavailable due to physical/ mental disability
3.11. Household has moved, new number available
3.12. Household has moved, new number not available

4. Refusal: (IF REFUSAL, ALSO ANSWER QUESTIONS 1 TO 4 BELOW)

4.1. Office refusal (i.e. in response to an advance letter, or in response to a message left on an answering machine or with someone else)

4.2. Proxy refusal by another person at the address (i.e. refuses to let interviewer speak with sample person)
4.3. Refusal by sample person at introduction/ before interview
4.4. Refusal during the interview (i.e. insufficient data to count as a partial interview)

5. Not eligible:

5.1. Out of sample (sample household does not belong to the survey population)
5.2. Sample household all deceased
5.3. Other

B. Call outcome codes for RDD samples (Germany, Hungary, Poland)

1. Contact, interview:

1.1. Complete Interview (for group C, this means both interviews complete)
1.2. Partial interview, i.e. break-off (for group C, this means complete part A and partial part B)

GROUP C ONLY:
1.3. Complete interview part A
1.4. Partial interview/break-off part A

For any call resulting in partial interview, interviewers should also record whether it is recommended to call again to complete the interview (e.g. circumstantial inability to continue the interview) or not (e.g. hard refusal). If refusal, also answer questions 1 to 4 below.

2. Non-contact:

2.1. Ring, no answer after 7 rings
2.2. Busy/engaged line
2.3. Answering machine/voice mail (residential); message left
2.4. Answering machine/voice mail (residential); no message left
2.5. Disconnected or other non-working
2.6. Temporarily disconnected
2.7. Fax/modem/data line/pager
2.8. Telecommunication technological barriers e.g. call barring, call screening
2.9. Technical phone problems e.g. bad line, Telephone Company experience technical problems
2.10. Call forwarded, no answer
2.11. Other non-contact - Give full details.

3. Contact – no interview:

3.1. Contact made at given phone number but selection procedure not completed; phone back at another time
3.2. Contact made, selection procedure completed but no contact with sample person
3.3. Contact made with sample person; appointment made
3.4. Contact made with sample person; phone back at another time
3.5. Contact made with sample person; other. Give details.
3.6. Sample person temporarily unavailable
3.7. Language barrier
3.8. Sample person unavailable (e.g. away, in hospital)
3.9. Sample person ill at home
3.10. Sample person unavailable due to physical/ mental disability

4. Refusal: (IF REFUSAL, ALSO ANSWER QUESTIONS 1 TO 4 BELOW)

4.1. Office refusal (i.e. in response to an advance letter, or in response to a message left on an answering machine or with someone else)
4.2. Proxy refusal by another person at the address (i.e. refuses to let interviewer speak with sample person)
4.3. Refusal by sample person at introduction/ before interview
4.4. Refusal during the interview (i.e. insufficient data to count as a partial interview)

5. Not eligible:

5.1. Out of service or disconnected
5.2. Non-residential (telephone number is used solely for businesses, schools and other organisations; does not include numbers that are shared for both business and private use).
5.3. Communal establishment/ institution (no private household(s))
5.4. Phone number is residential but no resident household (e.g. holiday homes)
5.5. Resident household(s) does not belong to survey population
5.6. Other

Annex 3. New questions added to the questionnaire

TF74 Can I just check, what kind of telephone are you using to talk to me? Is it … **READ OUT…**

…a fixed-line telephone with a wire attaching the handset to the base,	1	**GO TO TF76**
a fixed-line telephone with mobile handset,	2	
a mobile (cellular) phone,	3	**ASK TF75**
or something else? (WRITE IN: _____)	4	

TF75 Can I just check, are you at home at the moment or somewhere else? **(IF SOMEWHERE ELSE**: Where?)

At home	01
Someone else's home	02
At work/ office	03
In a car (driving)	04
In a car (not driving)	05
In a restaurant / bar	06
On a bus / train / tram	07
In a public place	08
Other (WRITE IN:)_____	09

ASK ALL

TF76 Many people find they are able to do other things while talking on the telephone, for example housework, watching television, reading or using a computer. During the course of the interview, were you doing anything else while we were talking? (**IF YES**: What were you doing?) **CODE ALL THAT APPLY**

No	1
Housework / cooking	2
Watching TV	3
Reading	4
Using a computer	5
Minding children	6
Other (WRITE IN:)_____	7

TF77 If you were asked to do a survey at home that would take about an hour, how would you choose to answer the questions? Would it be… **READ OUT…**

…face-to-face interview,	1
telephone interview,	2
filling in a paper questionnaire,	3
filling in a questionnaire on the web,	4
or, some other way? (WRITE IN) _____	5

TF78 Now we have finished the interview, I just want to ask you about the length of the interview. Would you have been willing to continue … **READ OUT…**

… much longer,	1
a bit longer,	2
or not at all?	3

Annex 4. Questionnaire adaptation.

A. Questions with short, simple response categories
Questions
B2, B3, B30-B33, B35-B37, C4, C7, C9, C13, C14, D20-D26, D40-D51, E4-E7, E23-E30, E40-E46, E53, E55, F5, F31, F33, F34
Change
Interviewers now read out the response categories.
Question stem has been amended and the instruction **'READ OUT…'** has been added.

B. Questions using response scales 0-10.
Questions
A8-A10, B4-B10, B23-B29, B34, B38-B40, C1, C21, D52-D54, E31-E39, E48-E51, E54, F18-F19, F46-F47
Change
Full description of scale and how to use it is given (if not already in the question). Interviewers now read out full question, including the description of the scale.

C. Questions with long and complex response categories.
Questions
A1-A6, C30, F6, F6a, F36, F49, F55
Change
Converted to open-ended questions. In some instances interviewer should code response, in others the response should be recorded verbatim.
Instructions have been added to signify what interviewers should do:

2 **'OPEN-ENDED AND CODE'** means interviewers should code respondents answer to the question using the list of response codes provided.
3 **'WRITE IN/ WRITE IN VERBATIM'** means interviewers should record responses in the spaces provided. Verbatim means word-for-word.

D. Miscellaneous
The following questions require more substantial changes:

1 A7, F62, F32 – these questions have been split into more than one question.
2 C2, C22, C23, E1, E3 – these questions now require the interviewer to 'read out' the response categories, and in addition some of the response categories have been collapsed.
3 E2 – this question required a slight wording change, has become a 'read out' question and some of the response categories have been collapsed.
4 E8-E22 – these questions are part of a long battery and although they have now become 'read out' questions the response categories should not be read out at every question, since this would become very repetitive. Some small wording changes are also needed to remind the respondent of the key aspects of the question stem.
5 F1 – this question requires a small wording change.
6 F8, F37 – these questions have changed substantially. Only 'main activity' is now collected and the question has been split into a number of sub-questions.
7 F54, F60 – these questions will now be post-coded by the interviewer, based on responses given to the previous questions.
8 F53b, F59b, F74-F78 – these are new questions that have been added to the telephone questionnaire. Version C additionally contains a new interviewer code in section X (X25).
9 F71 has been dropped.

E. No changes needed
(B1, B11-B22, C3, C5, C6, C8, C10-C12, C15-C20, C24-C29, C31-C36, D1-D19, D27-D39, D55, E47, E52, F2-F4, F7, F9-F17, F20-F30, F35, F38-F45, F48, F50-F53, F56-F59, F61, F63-F73)
No changes need to be made to these questions, other than removing references to showcards and small formatting changes in some cases.

Annex 5. Final outcomes by questionnaire version for each country.

Table A1 – Cyprus - Final outcomes by questionnaire version

	A		B		C		Total	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Issued Sample	420	-	226	-	410	-	1056	-
Ineligibles	20	4.8	17	7.5	22	5.4	59	5.6
Eligible sample	400	-	209	-	388	-	997	-
Full response rate (RR1)	34	8.5	11	5.3	20	5.2	65	6.5
Overall response rate (RR2)	35	8.8	13	6.2	35	9.0	83	8.3
Full cooperation rate (COOP1)	34	13.8	11	10.3	20	7.8	65	10.7
Overall coop rate (COOP2)	35	13.1	13	10.5	35	12.6	83	12.4
Contact rate (Household)	169	42.2	76	36.4	190	49.0	435	43.6
Contact rate (Respondent)	133	33.2	72	34.4	166	42.8	371	37.2
Refusal rate	122	30.5	52	24.9	144	37.1	318	31.9
Errors in contact data:								
- No outcome code	68	16.2[1]	27	11.9	52	12.7	147	13.9
- Duplicate codes	10	2.4	4	1.8	13	3.2	27	2.6

Notes: [1]% of total issued sample

Table A2 – Germany - Final outcomes by questionnaire version

	A		B		C		Total	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Issued Sample	618	-	309	-	618	-	1545	-
Ineligibles	13	2.1	5	1.6	7	1.2	25	1.6
Eligible sample	605	-	304	-	611	-	1520	-
Full response rate (RR1)	123	20.3	76	25.0	130	21.3	329	21.6
Overall response rate (RR2)	132	21.8	80	26.3	157	25.7	369	24.3
Full cooperation rate (COOP1)	123	24.7	76	28.8	130	25.5	329	25.9
Overall coop rate (COOP2)	132	26.6	80	30.3	157	30.8	369	29.1
Contact rate (Household)	497	82.1	264	86.8	509	83.3	1270	83.6
Contact rate (Respondent)	363	60.0	180	59.2	359	58.8	902	59.3
Refusal rate	352	58.2	177	58.2	331	54.2	860	56.6

Table A3 – Hungary - Final outcomes by questionnaire version

	A		B		C		Total	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Issued Sample	-	-	-	-	-	-	-	-
Ineligibles	-	-	-	-	-	-	-	-
Eligible sample	400	-	200	-	400	-	1000	-
Full response rate (RR1)	72	18.0	44	22.0	94	23.5	210	21.0
Overall response rate (RR2)	77	19.3	49	24.5	126	31.5	252	25.2
Full cooperation rate (COOP1)	72	19.9	44	25.6	94	25.7	210	23.4
Overall coop rate (COOP2)	77	21.3	49	28.5	126	34.4	252	28.0
Contact rate (Household)	361	90.2	172	86.0	366	91.5	899	89.9
Contact rate (Respondent)	342	85.5	161	80.5	354	88.5	857	85.7
Refusal rate	249	62.3	107	53.5	213	53.3	569	56.9

Table A4 – Poland - Final outcomes by questionnaire version

	A		B		C		Total	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Issued Sample	572	-	286	-	564	-	1422	-
Ineligibles	179	31.3	102	35.7	181	32.1	462	32.5
Eligible sample	393	-	184	-	383	-	960	-
Full response rate (RR1)	126	32.1	68	37.0	98	25.6	292	30.4
Overall response rate (RR2)	134	34.1	70	38.0	135	35.3	339	35.3
Full cooperation rate (COOP1)	126	42.4	68	54.0	98	36.7	292	42.3
Overall coop rate (COOP2)	134	45.1	70	55.6	135	50.6	339	49.1
Contact rate (Household)	297	75.6	126	68.5	267	69.7	690	71.9
Contact rate (Respondent)	171	43.5	85	46.2	162	42.3	418	43.5
Refusal rate	94	23.9	37	20.1	80	20.9	211	22.0

Table A5 – Switzerland - Final outcomes by questionnaire version

	A		B		C		Total	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Issued Sample	429	-	215	-	215	-	859	-
Ineligibles	23	5.4	5	2.3	6	2.8	34	4.0
Eligible sample	406	-	210	-	209	-	825	-
Full response rate (RR1)	154	37.9	83	39.5	56	26.8	293	35.5
Overall response rate (RR2)	154	37.9	83	39.5	105	50.2	342	41.5
Full cooperation rate (COOP1)	154	41.3	83	42.1	56	29.0	293	38.4
Overall coop rate (COOP2)	154	41.3	83	42.1	105	54.4	342	44.8
Contact rate (Household)	368	90.6	192	91.4	182	89.5	747	90.5
Contact rate (Respondent)	245	61.3	139	66.2	138	66.0	522	63.8
Refusal rate	129	31.8	68	32.4	50	23.9	247	29.9

Annex 6. Items included in each indicator of satisficing

Indicator	Type of scales	Item no.	No. of items
Non-differentiations	All scales	4-7 8-22 (-16) 23-30 35-39 40-45	4 14 8 5 6 **37**
Acquiesence	Agree/disagree scales	4-7 23-30 35-39	**18**
Mid-points	Agree/disagree scales	4-7 23-30 35-39	**18**
Primacy			
Strongly agree	Agree/disagree scales	4-7 23-30 35-39	**18**
Not at all	Not at all/a great deal scales	35-39	**5**
Recency			
Strongly disagree	Agree/disagree scales	4-7 23-30 35-39	**18**
A great deal	Not at all/a great deal scales	35-39	**5**