ISER Working Paper Series

Assessing the Effect of Data Collection Mode on Measurement

Annette Jäckle
Institute for Social and Economic Research
University of Essex

Caroline Roberts
Centre for Comparative Social Surveys
City University

Peter Lynn
Institute for Social and Economic Research
University of Essex

No. 2008-08
February 2008

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

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

www.iser.essex.ac.uk

Non-technical summary

Rising administration costs and falling response rates mean that many surveys that would previously have been carried out in one preferred mode of data collection are having to consider the use of mixed modes. For example, increasing numbers of surveys use a mix of modes, starting with a cheaper mode (such as telephone interviewing) which typically produces lower response rates, and following up non-respondents with face-to-face interviews. In order to decide about suitable data collection designs, survey practitioners must assess the trade-off between the potential advantages (for example in terms of financial costs and response rates) and disadvantages (for example in terms of data comparability) of mixing modes.

We discuss some of the challenges in evaluating the effects of using mixed modes on measurement and hence data comparability. The main argument is that it is very difficult to provide the information survey practitioners would need, about whether and to what extent using mixed modes would affect substantive conclusions. We briefly review theories about why different modes can lead to differences in survey responses. We then discuss the methods typically used to assess mode effects on measurement and then focus on some of the challenges. These include 1) the need to avoid confounding effects and what kinds of mode effects are actually identified, 2) the sensitivity of conclusions about the existence of mode effects to statistical methods used for the analysis of experimental mode comparison data, 3) the difficulty of assessing whether measurement differences matter in practice, and 4) the assessment of which mode provides better measurement. The main focus of the paper is on analysis methods. The points raised for discussion here arose in the context of the European Social Survey (ESS), which is conducting a programme of experimental research to inform the decision about whether to allow telephone interviewing in addition to face-to-face in its future rounds. We use some examples from the ESS experiments to illustrate how we tried to deal with these issues and to stimulate discussion. The paper concludes with an outlook of how the findings from the experimental studies are informing the decision process about whether or not to mix modes of data collection on the ESS and with general implications for mixed modes research.

Assessing the Effect of Data Collection Mode on Measurement

Annette Jäckle, Caroline Roberts and Peter Lynn

February 2008

ABSTRACT

We review the methods typically used to assess the effects of mode on measurement and data comparability and then discuss some of the challenges, including 1) the need to avoid confounding effects, 2) the sensitivity of conclusions to methods of analysing experimental mode comparison data, 3) the difficulty of assessing whether measurement differences matter in practice, and 4) the assessment of which mode provides better measurement. We illustrate the challenges and implications of mixed modes research for survey design with examples from experiments conducted in the context of the European Social Survey (ESS). The paper concludes with implications for mixed modes research.

Keywords: mixed modes, data comparability, analysis methods

JEL codes: C42

Acknowledgement:
The study reported derives from a programme of methodological research on mixed mode data collection co-funded by the European Commission as part of the central funding of the European Social Survey and Gallup Europe. The authors wish to thank Robert Manchin and Agnes Illyes, who directed the fieldwork at Gallup Europe; Roger Jowell, director of the European Social Survey; Norman Bradburn, Robert Groves, Willem Saris and members of the ESS Methods Group for advice and comments on the design of the experiment; Edith de Leeuw and Jop Hoox for detailed comments on an earlier version of the paper.

Contact:
Annette Jäckle, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Email: aejack@essex.ac.uk; Caroline Roberts, Centre for Comparative Social Surveys, City University, Northampton Square,London, EC1V 0HB, UK. Email: c.e.roberts@city.ac.uk; Peter Lynn, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Email: plynn@essex.ac.uk.

1 Introduction

Ever rising administration costs and falling response rates mean that many surveys are having to consider the use of mixed modes of data collection. Mixing modes offers the possibility of offsetting the disadvantages of one mode with the advantages offered by another, for example, combining cheaper modes with modes that lead to smaller survey errors (Dillman, 2000; de Leeuw, 2005). Mixing modes may however reduce data comparability, since different modes 1) provide access to different types of people, 2) attract different types of respondents, and 3) elicit different responses. As a result, the nature and magnitude of coverage, non-response and measurement errors may differ across modes, reducing the comparability of data collected with mixed modes.[1] The decision to mix modes therefore entails "an explicit trade off of errors and costs" (de Leeuw, 2005, p.235).

In order to make informed decisions, survey practitioners need to be able to evaluate and quantify the impact of mode on data quality. They also need to understand the causes of mode effects, if they are to design mixed modes studies that minimise the potential negative impact of mode on data comparability. These information needs are addressed by mode comparison studies, which typically assess the effect of mode on measurement either by testing the *comparability* of data collected in different modes (e.g. Hawkins, Albaum and Best, 1974; Groves and Kahn, 1979; Greenfield, Midanik and Rogers, 2000), or by testing specific hypotheses about the *potential causes* of differences between modes (e.g. Jordan, Marcus and Reeder, 1980; Holbrook et al, 2003; Fricker et al., 2005).

This paper is concerned with the first type of mode study and discusses some of the difficulties in evaluating whether mixing modes affects measurement and hence data comparability. In particular, we discuss the challenges in deriving information about whether mode effects would matter in practice that could feed into the decisions about suitable data collection designs.

The points raised for discussion here arose in the context of the European Social Survey (ESS), a biennial cross-sectional survey, that currently only allows face-to-face interviewing, but is considering the demands from several countries to allow alternative or complementary modes. To inform decisions about introducing mixed mode data collection on the ESS, the survey's coordinating team is carrying out a programme of experimental research designed to

[1] In their discussions of mixed modes designs, de Leeuw (2005) and Dillman (2000) distinguish between different stages of the survey, including the recruitment, presentation of questions and response and follow-up of non-respondents. In this paper we focus on the modes in which survey questions are presented.

assess the likely impact a move to other modes would have on data quality (see Jäckle, Roberts and Lynn, 2006 for details). When trying to interpret our results, we felt that it was not clear how findings about differences in responses translate into answers to the question "what would be the likely impact of a move to telephone interviewing?" It seemed hard to bridge the gap between 1) the kind of research that is feasible and typically done to evaluate mode effects, and 2) the information survey designers would actually need, about whether and to what extent mode effects would matter in practice. This paper discusses some of the difficulties in bridging this gap.

As a background to the discussion, we first summarise the theoretical literature describing how differences between modes of data collection can lead to differences in responses (Section 2) and describe the methods typically used to assess the impact of mixing modes on data quality (Section 3). We then briefly describe the ESS mixed modes experiments (Section 4), before discussing some of the challenges involved in attempting to evaluate the effects of mode. The first challenge we discuss is the need to avoid confounding factors in order to identify any effects of mode (Section 5.1). The main focus is then on analysis methods – how to analyse experimental mode comparison data in order to detect and evaluate mode effects (Section 5.2) and how to assess the magnitude of mode effects and their impact on data comparability (Section 5.3). The final challenge we discuss is how to assess the direction of mode effects and which mode provides better quality responses (Section 5.4). To illustrate, and stimulate discussion, we provide some examples of how we tried to deal with these various issues in our analysis of the ESS data. We then describe how the information from the experiments is informing decisions about whether and how to mix modes in the ESS (Section 6) and conclude with implications for the field of mixed modes research generally (Section 7).

2 How does mode affect measurement?

Cognitive models of the survey response process (e.g. Cannell, Miller and Oksenberg, 1981; Tourangeau, Rips and Rasinski, 2000) provide a useful framework for understanding how mode affects measurement. Differences in responses may arise if the process by which respondents come up with an answer is different in different modes. Characteristics of the mode may affect how the respondent understands the response task, retrieves relevant information, makes a judgement about the adequate response (which involves assessing the retrieved information and computing a response) and chooses the answer to report (see Roberts, 2007 for a discussion). As a result, differences between modes can lead to

differences in response biases, including a range of 'satisficing effects' and social desirability bias.

The theory of satisficing (Krosnick, 1991) posits that whether or not the respondent executes the response process optimally, or shortcuts instead, depends on the interaction between the difficulty of the task, the respondent's ability and motivation. Satisficing can, for example, be visible in the form of acquiescence (the indiscriminate use of 'yes' or 'agree' responses), non-differentiation (the indiscriminate use of one point on a response scale for a range of different items) or incomplete responses. Differences between modes may affect the amount of effort needed to answer the survey question or the respondent's motivation to make the required effort, leading to differences in the extent of satisficing between modes.

Social desirability bias (see DeMaio, 1984 for an overview) arises where respondents – either deliberately or unconsciously – select the more socially desirable response in order to portray themselves in a more favourable light than revealing the true answer would achieve. The respondent's willingness to report their answers accurately and honestly have been shown to be influenced by the perceived privacy of the survey setting, the perceived legitimacy of the survey and rapport between the interviewer and respondent (see Holbrook, Green and Krosnick, 2003). Differences between modes in these aspects are likely to lead to differences in the extent of social desirability bias between modes. In practice, the precise predictions as to how mode will affect respondents' reports, depends on the type of question, as well as on the type of mode (Biemer, 1988; Roberts, 2007).

3 Methods used to assess mode effects

Previous studies have typically compared the characteristics of data collected with different modes by testing for differences in a number of quality indicators and in response distributions (see, for example, Hawkins, Albaum and Best, 1974; Groves and Kahn, 1979; Greenfield, Midanik and Rogers, 2000). The quality indicators examined (see de Leeuw and van der Zouwen, 1988) include indicators of *completeness*, such as the mean item non-response rate across respondents in each sample, the mean length of responses to open-ended questions and the mean number of responses in 'tick all that apply' questions; indicators of response *accuracy,* such as comparisons with external data; and indicators of *reliability*, such as psychometric scaling properties. The next step is then typically to test for differences in the *response distributions* of the items under study. Conclusions about mode effects are then often drawn based on the proportion of indicators or items displaying significant differences across modes.

As Deming (1944, p.362) argued, however, "The problem is not whether differences [between modes] exist but how great are the differences, and why do they exist, and what effect will they have on the uses that are made of the data?". Similarly, Biemer pointed out that "statistical significance of a comparison alone is not necessarily indicative of a data quality differential between two surveys" (1988, p.276). He suggests three other factors that must also be considered: the effect size, the direction of the difference and potential confounding factors which could explain the mode differences. Starting from this framework we discuss some of the difficulties in assessing the impact of modes, using examples from the ESS mode comparison study for illustration.

4 The ESS mode experiments

The Central Co-Ordinating Team of the ESS is carrying out a programme of research investigating the feasibility of changing the current ESS policy of single-mode data collection using face-to-face interviews to a mixed mode data collection strategy in its future rounds. Part of this research has been conducted in collaboration with the Gallup Organisation, Europe. The two phases of the research described here formed part of this joint project (see Jäckle, Roberts and Lynn, 2006 for details).

Data collection for Phase I took place in May and June 2003 in Hungary. The study consisted of a 'hall test', in which participants, selected by quota sample to be representative of the Hungarian urban population by age, gender and education, were randomly assigned to one of four interview conditions: face-to-face interview, telephone interview, self-completion paper and pencil questionnaire and web-based questionnaire. Participants were then re-interviewed in a different mode[2]. All participants received the same questions in each of the four interviewing modes, making it possible to compare responses to different types of survey question across pairs of modes, and to examine differences in responses both between and within participants. Analysis of the phase I experiment (Peytcheva et al. 2004) indicated a number of areas that merited closer attention and the design for the second phase was drawn up in light of its conclusions.

Phase II involved a direct comparison between the current face-to-face methods employed on the ESS and telephone alternatives. Two experiments with the same design were conducted in Hungary and Portugal starting in July 2005. The experiments consisted of interviews conducted face-to-face in respondents' homes and telephone interviews conducted

[2] The experimental design was not fully balanced, however, and those interviewed by web in the first wave of data collection were not re-interviewed.

by fixed-line telephone (also in respondents' homes) or by mobile phone. The interviews consisted of a subset of questions from the core questionnaire of the ESS. In order to reduce costs, the fieldwork was concentrated in the countries' capital cities (Budapest and Lisbon), which also offered the advantage of suitable sampling frames in both locations, including telephone numbers and addresses. Each sampled address was randomly allocated to one of three treatment groups. At each contacted household, one person aged 15 or over was randomly selected for interview using the last birthday method. The examples used here are from the experiment carried out in Hungary (but excluding the mobile phone group), where 515 respondents were interviewed face-to-face using showcards, 518 respondents were interviewed face-to-face without showcards and 887 were interviewed over a fixed-line telephone using the same questionnaire administered to the face-to-face-without-showcards group.

5 Challenges in assessing mode effects

5.1 Confounding factors in studies of mode effects

Mode comparison research has mainly consisted of two types of studies: those comparing systems of data collection (Biemer, 1988; de Leeuw, 2005) - where an optimal design in one mode is compared with an optimal design in another mode - and studies attempting to identify a so-called 'pure' mode effect. System comparisons provide realistic information about the effects of different modes on data comparability. Since each system will differ in many respects, for example, in terms of coverage, sampling frame, non-response bias and questionnaire (see Holbrook et al., 2003), results from such studies cannot be generalised, even to other comparisons of the same modes, and cannot be used (on their own) to predict mode effects in other settings. In contrast, the second type of study has used experimental designs to attempt to isolate the effect of mode on measurement, controlling for other characteristics of the survey (though in practice few are successful; see Holbrook et al., 2003). Such studies may further attempt to isolate the effects of different characteristics of modes, such as the effects of using showcards, or of the presence of the interviewer in face-to-face interviews (see Example 3).

The first challenge is that identifying the net effect of mode on measurement requires careful experimental designs, such that the only difference between samples is the mode they are assigned to. Only under such circumstances can differences in responses be attributed to the mode; if any other aspect differs between the samples, the effect of mode is confounded with these other differences.

Example 1: To hold any error from sampling and coverage consistent across the experimental groups, the Phase II ESS experiment used telephone listings containing address information as frames for both the face-to-face and telephone samples. An equal-probability sample of fixed residential phone numbers within the defined area (the Greater Budapest region of Hungary) was selected and each sample unit was then randomly allocated to one of the mode treatments.

Example 2: The problem of differential nonresponse across treatment groups can be avoided by randomly allocating respondents to mode *after* the recruitment. The Phase I ESS experiment, for example, involved a hall test, in which respondents were randomly allocated to face-to-face, telephone, web and self-completion once they had agreed to participate. With such a design any differences in responses can clearly be attributed to the mode. This however comes at the cost of some realism, since 'real-world' survey interviews are typically conducted under quite different settings, in respondents' homes, with family members present, etc.

Differences in the questionnaire can further confound the effect of mode. According to Dillman (2000) one of the biggest causes of apparent mode effects is the tendency for questions to be constructed differently for different types of questionnaires. In the case of telephone and face-to-face interviewing, a further potential confounding difference is the use of showcards: any observed differences in responses could be attributed either to differences in the use of visual aids – or to other differences between the modes, such as the social distance when the interviewer is separated from the respondent by telephone. To avoid this pitfall, Dillman advocates the 'unimode' construction approach to questionnaire design for multimode surveys, in which questions are designed from the outset to be suitable for administration in all modes. However, in practice, many survey designers considering mixing modes of data collection are doing so in the context of an existing survey and questionnaire design for the survey is constrained by the design of questions in the primary mode.

Example 3: In the Phase II ESS experiment, the differences in responses due to the use of showcards and differences due to other mode differences were isolated by including three treatment groups: 1) face-to-face with showcards, 2) face-to-face without showcards, and 3) telephone using the same questionnaire as group 2). This design allows the identification of different types of mode effects: the net mode effect between the current ESS *system of data collection* (using showcards) and telephone can be identified by comparison of groups 1) and 3); the *showcard effect* can be identified by comparison of groups 1) and 2) and the effect of

residual mode differences, including the physical presence of the interviewer, can be identified by comparison of groups 2) and 3).

Differences due to potential differential non-response however remained a problem. The response rates were comparable in both samples, but low (32% in telephone and 33% in face-to- face). The telephone sample had a significantly lower proportion of men, manual workers and respondents with low education levels. There were however no differences across modes in mean age and the proportion in work. This is consistent with the findings of other studies (see Holbrook et al., 2003).

5.2 Methods of assessing mode effects

The second challenge in mode research stems from the problem that if the sample composition differs between modes because of differential non-response, then any tests for mode effects need to control for respondent characteristics. One option is to use weighting methods, whereby observations in one sample are weighted such that they reflect the characteristics of the 'reference' sample (see for example Fricker et al., 2005). The weighted samples can then be used to calculate significance tests, such as t-tests of the equality of sample means or Chi-square tests of whether response distributions are independent of mode. An alternative is to use different regression approaches, depending on the level of measurement of the item. As the examples below show, for ordinal items the choice of regression model can affect conclusions about the existence of mode effects.

For continuous dependent variables OLS regressions can be used, including the sample characteristics and a mode dummy as explanatory variables. A t-test can then be performed to test the null hypothesis that the mode coefficient is not significantly different from zero. In the case of nominal categorical variables multinomial models can be used, which fit a different model for each of the response categories. For ordinal categorical variables, such as the attitudinal indicators carried in the ESS, OLS estimation might provide misleading results, since the intervals between adjacent response categories cannot be assumed to be equal. Compared to the multinomial model, the additional information contained in the ordering can, however, be used to estimate more parsimonious models. In general, models for ordinal variables assume that higher values of the dependent variable correspond to higher outcomes, but the actual values are irrelevant.

The proportional odds modelling technique (also referred to as cumulative odds model (O'Connell, 2006), parallel regression model, or grouped continuous model (Long, 1997) is, according to Billiet and Welkenhuysen-Gybels (2004) currently the best method available for

assessing measurement equivalence of ordinal data. The proportional odds model is equivalent to a sequential series of binary logistic regressions of P($Y>j$) over cumulative splits of the data, where the coefficients are constrained to be equal in each equation:

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta)}{1+[\exp(\alpha_j + X_i\beta)]} \quad j = 1, 2, ..., M-1$$

where M is the number of response categories of the ordinal dependent variable Y. For a variable with 4 response categories there are 3 cumulative dichotomisations j = 1, 2, 3. In this case, the proportional odds model is equivalent to estimating the sequence of binary logistic regressions of $P(Y_i>1)$, $P(Y_i>2)$ and $P(Y_i>3)$, for the different cumulative dichotomisations of the response categories (see O'Connell (2006) for ordinal models representing alternative dichotomisations. Compared to separate estimations of each cumulative dichotomisation, the results differ slightly when all equations are estimated simultaneously (Williams 2006)). The proportional odds model constrains the coefficients to be equal in each of the cumulative splits, assuming that the odds(j) = P($Y \leq j$) / P($Y>j$) have the same ratio for all combinations of explanatory variables for any dichotomisation j. That is, the model assumes that covariates 'shift' the distribution of responses proportionately across all categories. In the case of mode effects, the model would assume that if telephone respondents say they are more interested in politics than face-to-face respondents, then this shift should be visible in all response categories. This assumption of proportional odds may however not hold. In fact, most theories of the causes of mode effects posit that responses differ because some categories (for example 'agree' in a scale from 'strongly agree' to 'strongly disagree') are disproportionately selected, implying that mode effects are not necessarily proportional across response categories.

The assumption that mode has a proportional effect on all response categories can be tested using *partial* proportional odds models. In this case explanatory variables for which the proportional odds assumption holds are constrained to be equal, while variables for which the assumption is violated are allowed to vary across the cumulative dichotomisations. This model has been implemented in Stata by Williams (2006). The estimation uses a backwards stepwise selection procedure, starting with a model corresponding to the full set of cumulative logistic models, where a different set of coefficients is estimated for each model, and gradually imposing constraints for variables for which the assumption of proportional odds holds, based on Wald tests of the equivalence of coefficients across equations. For items where the proportional odds assumption holds, the standard error of the single mode

coefficient can be used to test for mode effects on the response distribution. For items where the proportional odds assumption does not hold, one can test for the overall effect of mode using Wald tests of the joint hypothesis that all mode coefficients from the series of cumulative logistic dichotomisations equal zero, as well as reporting the significance of mode coefficients from each dichotomisation.

Example 4: In the Phase II ESS experiment, we tested all ordinal variables for mode effects using 1) partial proportional odds models, 2) proportional odds models (ologit) and 3) ordinary linear models (OLS). For each ordinal question, the response was predicted by the mode indicator and controls for the socio-demographic composition of the samples, including age, age squared, sex, educational qualification and occupation. The results suggest that the analysis method may well affect conclusions about the existence of mode effects. The examples presented here tested for differences between the face-to-face showcard and the telephone groups, identifying the 'system effect' discussed above. The results are only shown for those variables with a significant mode effect. For information about the 12 ordinal variables, which did not show any mode effects, and the results of the other two treatment comparisons, identifying the 'showcard effect' and the 'interviewer effect', see Jäckle et al. (2006).

For each question, Tables 1 and 2 present the predicted response distributions in the face-to-face mode. The predictions are reported for both the partial proportional odds and the ologit estimations. The tables then present the percentage point differences in predicted response probabilities due to mode. The P-values indicate the significance level of the mode coefficient. In Table 2, two types of P-values are presented for the proportional odds model: for each item the first row is the P-value from a joint Wald test that the mode coefficients for all dichotomisations are zero. The remainder rows show the P-values of the mode coefficient for each dichotomisation. For example, the P-value corresponding to the third response category corresponds to the mode coefficient of the model $P(Y_i>3)$. The results show that the proportional odds assumption holds for some items (Table 1), but not for all (Table 2).

For the items where the proportional odds assumption holds, similar conclusions about the mode effect would have been drawn based on the ologit and OLS models: the significance levels of the mode indicator are similar in all three models (except for Q6) and the differences between modes in response distributions are similar in the partial proportional odds and ologit models. Telephone respondents were:

- more likely to be 'very' or 'quite' interested in politics (Q5) and less likely to find politics too complicated (Q6);

- more likely to support allowing ‘a few’, ‘some’ or ‘many’ immigrants from poor countries outside Europe to live in the country (Q13);
- more likely to say immigration enriches cultural life (Q15) and less likely to say it makes their country a worse place to live (Q16);
- less likely to (strongly) agree that mothers should cut down on paid work (Q17a) and that men have more right to jobs than women when jobs are scarce (Q17c);
- more likely to strongly agree that men should share responsibilities for their home and family (Q17b) and that the law should be obeyed whatever the circumstances (Q18b).

For three of the seven items for which the proportional odds assumption did not hold, the ologit and OLS models would have lead to the conclusion that there was no significant mode effect. For these variables the predicted response distributions based on the partial proportional odds model illustrate the non-linearity of the mode effect. Telephone respondents were:

- more likely to choose 0, 5 or 8 on an 11 point scale of their level of trust in the EU parliament (Q8f);
- less likely to choose 1 or 2 and more likely to choose 5 or 6 on an 11 point scale of how good immigration is for the economy (Q14);
- less likely to strongly agree or strongly disagree that parents should stay together even if they do not get along (Q17d).

The differences in response probabilities predicted by the ologit model do not capture these non-linearities. Imposing the proportional odds smoothness assumption produces predicted response distributions that hardly differ between the two modes. That is, using the ologit model analysts would falsely conclude that mode does not affect measurement, when in fact there are mode effects.

For the remaining four items the ologit and OLS models did capture the mode effect, estimating similar significance levels for the mode coefficient as for the joint Wald test from the partial proportional odds model. Nonetheless, the predicted probabilities again show that imposing the ologit smoothness assumption masks the non-linearities of the mode effects, producing different signs and magnitudes of differences between modes for some of the response categories. For example, telephone respondent were:

- much more likely to report watching TV for anywhere between ½ an hour and 1½ hours and much less likely to report watching for more than 2½ hours (Q1). The

ologit model predicted much smaller differences between modes for these two intervals;

- much less likely to report watching news on TV for less than ½ hour and much more likely to report watching for anywhere between ½ an hour and 1 hour (Q2); the ologit model predicted less difference between modes for the first of these intervals, and hardly any difference for the second;
- less likely to be against allowing any immigrants of the same ethnicity (Q11) or of a different ethnicity (Q12), but in both cases more likely to want to allow 'a few', 'some' or 'many'. For these items the ologit model estimated that telephone respondents were also less likely to choose 'a few'.

An alternative method of testing for mode effects while controlling for sample composition would be to use structural equation modelling to assess equivalence of measurement. Billiet and Welkenhuysen-Gybels (2004) compared this approach with the proportional odds models discussed above and used both to assess equivalence of six immigration items (three of which were also carried in the present experiment) across 21 countries in the first wave of the ESS. The authors concluded that structural models lacked power and that the proportionate odds model was the currently best available method of testing for equivalence of Likert scale items.

Regardless of the analysis method used, testing for mode effects is typically a 'fishing' exercise, where all items carried in the questionnaire are tested for differences. (This may be different for studies testing hypotheses about the *causes* of mode effects, which only test specific indicators. For the Phase II ESS study, for example, we first tested for differences in the measurement of all items between modes (the focus of this paper), before testing specific hypotheses about the causes of mode effects (see, Roberts et al. 2006)). The researcher decides on the level of significance, say 0.05, implying that they are willing to accept a 5% chance of falsely accepting a rare occurrence, by chance or due to sampling error, as evidence of a mode effect, when in reality there is none. However, with multiple tests from a single experiment such as here, where we tested all 28 ordinal items carried in the questionnaire, the risk of accepting false positives as significant mode effects increases with every additional item tested. In the worst case, if we were testing 28 independent hypotheses, the risk of false positive inference would increase 28 times.

Example 5: To adjust for multiple comparisons, we used the Bonferroni-Holm method (as described in Ludbrook 1998) to adjust the P-values from the partial proportional odds

models.[3] With this method, the P-values are sorted and starting with the smallest, the adjusted value is calculated as $P_b = m \times P$, where m is the number of hypotheses (in this case items) tested. For the next smallest P-value the adjusted value is calculated as $P_b = (m-1) \times P$ and so forth until P_b>0.05. The raw P-values from the partial proportional odds models in Tables 1 and 2 suggested significant mode effects at the 0.05 level for 16 of the 28 items tested. After applying the Bonferroni-Holm adjustment, the adjusted P-Values for Q6, Q8f, Q13, Q17b and Q17c were larger than 0.05. That is, without the adjustment we might have concluded that mode affected more than half (57%) of ordinal items studied. After excluding what are likely to be false positives, we conclude that the responses to 39% of items were affected by mode.

5.3 Assessing the size of mode effects

Most studies assess the extent to which the mode of data collection affects measurement using tests of statistical significance to evaluate the size of differences in responses across modes. Few studies have, however, attempted to assess the significance of any observed effects in terms of whether and how they might affect the substantive conclusions drawn by data analysts. We would argue that in order to evaluate whether mode affects data comparability it is necessary to move away from an assessment of means and marginal distributions toward an assessment of the effect of mode on relevant estimates.

The third challenge is that differences in responses across modes may impact on certain types of estimates, but not on others. Conclusions about the effect of mode therefore depend on the application to which the data are put. De Leeuw (1992) similarly argued that the prevalent focus on effects on univariate estimates is not enough to fully evaluate the effects of mode on estimates. As well as testing for univariate mode effects in response styles (such as acquiescence or extremeness), she also tested for psychometric mode effects (that is the reliability and scalability of items collected with different modes) and multivariate mode effects (by replicating substantive applications of structural equation models).

Example 6: In the ESS data, the mean summed score on a set of items measuring attitudes toward immigration was 0.46 for face-to-face and 0.51 for telephone (P=0.001). (The attitude

[3] The reason for choosing the Bonferroni-Holm method over other methods was that it is less conservative than Bonferroni, especially when the multiple hypotheses are not independent, very simple to compute, and very versatile in that it can be used for continuous, ordinal and categorical data. Methods of adjusting for multiple comparisons are usually described for tests of means, but can also be used for regression coefficients (see, James 1964).

scores were calculated by summing up the responses to three 4-category items and three 11-category items and normalising the summed score to lie between 0 and 1. Higher values represent being in favour of immigration.) If different subgroups, or countries, are surveyed with different modes, then a mode effect of this kind may lead analysts to falsely conclude that there are differences between these groups. The primary usage of these data may, however, not be to test for differences in levels between countries, but for differences in the determinants of immigration attitudes. We therefore tested whether mode affected the relationship with other variables, using OLS models to regress the summed immigration score on mode, a predictor variable, its interaction with mode and controls for socio-demographic sample composition (Table 3). The predictor variables included a series of binary indicators (whether in work, whether voted in last elections, whether voted for a centre/right or socialist/liberal party and whether had access to internet), some 11-point scale items (trust in people, life satisfaction and religiosity) and the summed attitude scores for political interest and gender-role attitudes, both of which had different distributions across the modes. Respondents who had voted, had internet access, were more likely to trust other people or were more satisfied with their life tended to have higher (more liberal) immigration scores; respondents who had less of an interest in and understanding of politics had lower scores on the immigration scale. Mode did not affect the relationships with any of the nine predictors, except for voting, where the difference between voters and non-voters disappeared in telephone mode. This suggests that even if analysts were to reach different conclusions about the *levels* of immigration attitude scores for sample members surveyed with different modes, conclusions about *differences between subgroups* need not be affected.

A further challenge is that survey methodologists can really only uncover statistical differences between estimates. Whether these matter in practice, that is, whether differences affect conclusions, depends on the substantive interpretation of results by data users. Differences in estimates from different modes may be significant, but they need not be substantively important. In De Leeuw's (1992) analysis, for example, the estimated determinants of loneliness in a structural equation model were comparable across modes; their relative importance however differed. Conclusions about the impact of mode on data comparability will therefore differ depending on which aspect is the principal objective of the analysis.

5.4 Direction of mode effects

A further step in evaluating the effect of mode on data quality according to Biemer (1988) is to judge which mode provides better quality data. This can be done by comparing responses with a 'gold standard' such as external records, assessing internal validity of responses, or based on prior knowledge about the direction of errors, where for example higher reporting of sensitive behaviours is judged as better. This approach presumes a situation where the survey agency is free to choose the mode which performs best. In many situations, in which decisions about mixing modes are made, however, there is a pre-specified primary mode, which in the case of a panel or repeated cross-sectional study may already have been implemented in prior surveys. The question then is not which mode produces least response errors, but whether the errors in different modes are comparable (Braun, 2003). If errors are not comparable, then introducing mixed modes could not only lead to misleading conclusions about differences between sample members within a cross-section, but also about differences over time.

6 Implications of mixed modes research for survey design – the example of the ESS

The findings from the experimental studies encouraged the ESS team to consider the possibility of allowing telephone interviewing under certain conditions. To summarize the findings of the mode experiment, we found significant differences in responses between modes for 11 of the 28 ordinal items tested.[4] For 5 of these items the mode effect was linear, such that telephone respondents consistently reported higher or lower response categories; for 6 items the effects were non-linear, where mode affected the extremes or middle response categories disproportionately. At the same time, however, mode did not affect bivariate relationships between variables. The results therefore suggested that allowing telephone interviewing in addition to the current face-to-face interviewing could affect estimates of means and prevalences, but would not necessarily affect estimates of relationships between variables. We could however not resolve the issue of how to decide whether a significant difference in responses or estimates would matter in practice, since this would depend on the specific substantive interpretation. In addition, there is an indefinite number of applications for which these data may be used, which we did not, and cannot, assess. Any conclusions about whether allowing telephone mode effects in a mixed mode ESS would matter in

[4] The items included in the experiment were a subset of the ESS core questionnaire, that were expected to be most susceptible to mode effects. We would, therefore, expect the overall proportion of items in the full ESS questionnaire susceptible to mode effects to be smaller.

practice and affect substantive conclusions drawn by analysts are therefore not generalisable across estimates.

Many questions also remain about how to develop equivalent questions for use in alternative modes that will preserve the continuity of estimates in the time series. In order to develop ways of reducing mode effects on measurement, the experiment was also designed to test some hypotheses about the causes of mode effects. The findings suggested that differences were mainly due to the presence of the interviewer; showcards did not appear to affect responses. The main difference appeared to be more social desirability bias in the telephone mode than face-to-face; however, there was no evidence that telephone respondents were more likely to satisfice. A major limitation of the experimental study was, however, that it used a considerably abridged version of the ESS questionnaire, which is unlikely to have provided an adequate test of our hypotheses concerning the increased likelihood of satisficing in telephone interviews. Only by conducting long interviews by telephone can we establish whether or not satisficing effects are likely to be detrimental to data quality. Partly for this reason – and partly to explore the difficulty in itself of administering by telephone a long survey questionnaire like that used in the ESS – the most recent phase of the ESS mixed mode research has conducted precisely such a test.

Finally, there are other elements that would need to go into the cost-benefit analysis of allowing telephone interviewing, which are still unknown. The ESS experiments so far have focused on the effects of mixed modes on measurement. The effects on coverage and non-response errors would also need to be evaluated, since these also impact on data comparability. Regarding benefits, comparatively little research has empirically established whether the apparent advantages of a multi-mode data collection strategy – in terms of cost-savings or a possible increase in response rates - would be proven in reality (exceptions include Dillman et al. 2001; Hochstim, 1967; Mooney, Giesbrecht and Shettle, 1993; and Voogt and Saris, 2005). Similarly, while there is much anecdotal evidence about the apparent need to tailor data collection designs to the different survey climates of the countries participating in the ESS, until recently, relatively little was known about the actual demand for alternatives to face-to-face interviewing and the capacity for adopting different approaches. For these reasons, the ESS team has also carried out a 'mapping exercise' to build up a portrait of current survey practice across Europe and to establish a basis for thinking about how to design the most suitable multi-mode data collection strategy for a cross-national time-series (see Roberts, Eva and Widdop, 2008).

7 Implications for mixed modes research

Although there are literally hundreds of studies that have tested the comparability of data collected with different modes, we are still faced with large uncertainties when making decisions about survey designs. In this paper, we discussed some of the difficulties in assessing the effect of mode on measurement and producing information that could inform decisions about the trade-offs in using mixed modes.

It is extremely difficult to devise mode comparisons such that any differences in responses can clearly be attributed to the effect of mode on measurement. In particular, mode effects are often confounded with differences in sample composition due to differences in coverage, sampling error or non-response bias associated with different modes. As a result, even experimental comparisons need to be analysed with sophisticated statistical methods to take into account, at least some of, the differences in sample composition. As the examples in this paper have highlighted, choosing appropriate statistical methods is important and can affect conclusions about the existence of mode effects. This is in particular complicated by the non-linear nature of many mode effects, which implies that mode does not simply cause a shift in the distribution of responses, but can lead to more complicated differences between modes. A further implication of such non-linear mode effects, which in our view is not sufficiently recognised, is that it is not enough for analysts to simply 'control for mode', for example, by including a dummy variable for mode in a regression. The question what would be appropriate adjustment methods for non-linear mode effects is however still unresolved. In any case, a prerequisite for choosing an appropriate adjustment would be the ability to predict the likely nature of the mode effect for any given item. At present, we still lack underlying theories that would allow us to do so. In the findings presented here it is, for example, not obvious why some items had linear and other related items had non-linear mode effects.

Even if appropriate methods are used to test for differences in responses which might be attributed to mode, this does not answer the question whether these differences would matter in practice. The difficulty is that a given difference in responses to survey questions may cause biases in some types of estimates, but not in others. In our view, to answer the question that ultimately motivates all research in this field, which is whether mixing modes would affect substantive conclusions, we need to rethink the types of tests we perform in order to draw conclusions about data comparability. More informative tests would be motivated by the applications to which the data are put, for example replicating existing research using data collected in each mode. This may require (more) collaboration between survey methodologists and substantive researchers and may help us identify conditions under

which mode effects on measurement matter or do not matter. We would note that such tests are likely to be more difficult to identify in the case of multi-purpose public-use datasets, compared to surveys with a specific focus.

Finally, the effects of mode on measurement need to be evaluated in the context of the effect of mode on other survey errors. Janssen and Schouten (2007) have suggested that such a total survey error perspective could, for example, be created by developing quality indicators of the effects of mode on different error types.

References

Biemer, P. P. 1988. "Measuring Data Quality." Pp. 273-282 in *Telephone Survey Methodology*, edited by R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, and J. Waksberg. New York: Wiley.

Billiet, J. and Welkenhusen-Gybels, H. 2004. "Assessing Cross-National Construct Equivalence on the ESS: The Case of Six Immigration Items." In *Recent Developments and Applications in Social Research Methodology,* edited by C. Van Dijkum, J. Blasius, and C. Durand. Barbara Budrich Pub. Proceedings of the Sixth International Conference on Social Science Methodology. Amsterdam.

Braun, M. 2003. "Errors in Comparative Survey Research: An Overview." In *Cross-Cultural Survey Methods*, edited by J. A. Harkness, F. J. R. Van der Vijver and P. P. Mohler. Hoboken, NJ: John Wiley.

Cannell, C., Miller, P., and Oksenberg, L. 1981. „Research on Interviewing Techniques." Pp. 389-437 in *Sociological Methodology 1981*, edited by S. Leinhardt. San Francisco: Jossey-Bass.

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

–. 2005. "To Mix or not to Mix? Data Collection Modes in Surveys." *Journal of Official Statistics* 21:1-23.

de Leeuw, E. and van der Zouwen, J. 1988. "Data Quality in Telephone and Face-to-Face Surveys: A Comparative Analysis." Pp. 283-299 in *Telephone Survey Methodology*, edited by R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II and J. Waksberg. New York: Wiley.

DeMaio, T. J. 1984. "Social Desirability and Survey Measurement: A review." Pp. 257-282 in *Surveying Subjective Phenomena*, edited by C. F. Turner and E. Martin. New York: Russell Sage Foundation.

Deming, W.E. 1944. "On Errors in Surveys." *American Sociological Review* 9:359-369.

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

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., and Berck, J. 2001. "Response Rate and Measurement Differences in Mixed Mode Surveys using Mail, Telephone, Interactive Voice Response and the Internet." *Unpublished paper.*

http://www.sesrc.wsu.edu/dillman/papers/Mixed%20Mode%20ppr%20_with%20Gallup_%20POQ.pdf

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69:370-392.

Greenfield, T. K., Midanik, L. T., and Rogers, J. D. 2000. "Effects of Telephone versus Face-to-Face Interview Modes on Reports of Alcohol Consumption." *Addiction* 95:277-284.

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

Hawkins, D. I., Albaum, G., and Best, R. 1979. "Stapel Scale or Semantic Differentail in Marketing Research?" *Journal of Marketing Research* 11:318-322.

Hochstim, J. R. 1967. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association* 62:976-989.

Holbrook, A. L., Green, M. C., and 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.

Janssen, B. and Schouten, B. 2007. "Survey Errors in Mixed Mode Pilots that Incorporate Face-to-Face, Telephone, Paper and Web Interviewing." Presented at the European Survey Research Association Conference, Prague.

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

Jäckle, A., Roberts, C. E., and Lynn, P. 2006. "Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes (Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project)." *ISER Working Paper,* 2006-41. Colchester: University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2006-41.pdf

Krosnick, J. A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213-236.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Ludbrook, J. 1998. "Multiple Comparison Procedures Updated." *Clinical and Experimental Pharmacology and Physiology* 25:1032-1037.

Mooney, G., Giesbrecht, L., and Shettle, C. 1993. "*To Pay or not to Pay: That is the Question."* Presented at the meeting of the Amercian Association for Public Opinion Research, St. Petersburg, FL.

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

Peytcheva, E. A., Manchin, R., Tortora, R., and Groves, R. M. 2004. "Comparing Face to Face, Telephone, Paper Self-Administered and Web Survey Measurement." Presented at the American Association for Public Opinion Research 59th Annual Conference, Phoenix, Arizona.

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

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." *Working Paper of the Centre for Comparative Social Surveys,* 2008-01. London: City University.

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/

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

Voogt, R. J. J. and Saris, W. E. 2005. "Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21:367-387.

Williams, R. 2006. "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables." *The Stata Journal* 6:58-82.

Table 1: Mode effects for which proportional odds assumption holds

Question	Response categories	Partial Proportional Odds			Proportional Odds (Ologit)			OLS
		F2F (Col %)	Tel-F2F (% Pts)	P-Value[1]	F2F (Col %)	Tel-F2F (% Pts)	P-Value	P-Value
Q5 Political interest?	Very interested	12.60	5.87	0.000	12.40	6.30	0.000	0.000
	Quite interested	39.63	4.99		39.19	5.24		
	Hardly interested	29.34	-4.78		29.76	-5.32		
	Not at all interested	18.42	-6.06		18.65	-6.22		
Q6 Politics too complicated?	Never	19.42	4.36	(0.033)	18.99	4.68	ns	ns
	Seldom	21.30	2.27		21.28	2.47		
	Occasionally	32.44	-1.34		32.20	-1.31		
	Regularly	11.35	-1.75		11.52	-1.88		
	Frequently	15.49	-3.54		16.01	-3.96		
Q13 Immigration: poor outside Europe?	Allow many	6.07	1.90	(0.046)	6.13	2.05	0.012	0.017
	Allow some	16.21	3.52		16.38	3.63		
	Allow a few	46.05	0.85		46.54	0.29		
	Allow none	31.68	-6.29		30.96	-5.97		
Q15 Immigration: impact on culture?	0 Cultural life undermined	7.10	-2.56	0.001	6.86	-2.24	0.000	0.000
	1	3.15	-0.99		3.10	-0.95		
	2	4.69	-1.15		4.81	-1.38		
	3	6.72	-1.70		6.60	-1.71		
	4	6.05	-1.26		6.33	-1.42		
	5	23.99	-2.96		23.94	-3.01		
	6	9.00	0.25		9.20	0.01		
	7	10.11	0.86		10.09	0.92		
	8	15.21	3.66		15.33	3.60		
	9	3.95	1.41		3.97	1.45		
	10 Cultural life enriched	10.03	4.44		9.78	4.72		
Q16 Immigration: impact on living standards?	0 Worse place to live	12.71	-4.89	0.000	12.64	-5.17	0.000	0.000
	1	3.52	-1.16		3.49	-1.21		
	2	7.94	-2.29		7.89	-2.38		
	3	12.05	-2.57		11.98	-2.64		
	4	7.28	-0.99		7.36	-1.03		
	5	35.58	2.44		35.53	2.52		
	6	6.87	2.23		7.12	2.39		
	7	5.17	2.20		5.18	2.29		
	8	5.75	3.05		5.79	3.22		
	9	0.79	0.47		0.76	0.48		
	10 Better place to live	2.34	1.50		2.26	1.53		
Q17a Gender role: mothers should not work?	Agree strongly	23.14	-6.63	0.001	23.07	-6.47	0.000	0.000
	Agree	29.77	-3.74		29.98	-3.79		
	Neither	20.21	1.46		20.41	1.42		
	Disagree	20.42	5.89		20.23	5.85		
	Disagree strongly	6.46	3.03		6.32	3.00		
Q17b Gender role: Men responsible for family?	Agree strongly	74.10	6.67	(0.006)	74.01	7.03	0.009	0.001
	Agree	20.51	-5.00		20.53	-5.26		
	Neither	3.03	-0.92		2.85	-0.90		
	Disagree	1.92	-0.60		2.18	-0.72		
	Disagree strongly	0.44	-0.14		0.44	-0.15		
Q17c Gender role: Men more right to jobs?	Agree strongly	27.14	-5.98	(0.014)	27.63	-6.58	0.021	0.016
	Agree	17.25	-1.65		17.32	-1.60		
	Neither	22.29	0.40		21.99	0.56		
	Disagree	23.96	4.16		23.82	4.35		
	Disagree strongly	9.36	3.07		9.25	3.27		
Q18b Law should always be obeyed?	Agree strongly	61.24	12.17	0.000	61.06	11.79	0.000	0.000
	Agree	23.04	-6.06		23.21	-5.94		
	Neither	11.46	-4.30		11.11	-3.98		
	Disagree	3.39	-1.44		3.67	-1.47		
	Disagree strongly	0.88	-0.39		0.95	-0.40		

[1] P-Values in brackets were larger than 0.05 after adjusting for multiple comparisons using the Holm-Bonferroni procedure.

Table 2: Mode effects for which proportional odds assumption does not hold

Question	Response categories	Partial Proportional Odds			Proportional Odds (Ologit)			OLS
		F2F (Col %)	Tel-F2F (% Pts)	P-Value[1]	F2F (Col %)	Tel-F2F (% Pts)	P-Value	P-Value
Ql				0.0000			0.004	0.002
Time watching TV?	0 hrs	2.59	0.22	ns	1.99	1.03		
	0 - 1/2 hr	5.43	-3.52	0.016	2.35	1.07		
	1/2 - 1 hr	12.73	13.14	0.000	18.28	5.03		
	1 - 1 1/2 hrs	7.42	1.24	0.000	7.74	1.02		
	1 1/2 - 2 hrs	21.04	-0.11	0.001	20.78	0.65		
	2 - 2 1/2 hrs	6.81	-0.43	0.001	6.78	-0.35		
	2 1/2 - 3 hrs	18.80	-6.78	ns	15.23	-1.75		
	> 3 hrs	25.17	-3.75	–	26.85	-6.70		
Q2				0.0000			0.000	0.001
Time watching TV news?	0 hrs	4.62	-0.10	ns	6.74	-2.87		
	0 - 1/2 hr	32.98	-22.69	0.000	23.89	-8.26		
	1/2 - 1 hr	36.99	21.50	ns	51.83	-0.61		
	1 - 1 1/2 hrs	11.19	-2.03	ns	7.13	3.46		
	1 1/2 - 2 hrs	7.26	3.98	ns	6.52	4.63		
	2 - 2 1/2 hrs	1.53	0.13	ns	0.99	0.86		
	2 1/2 - 3 hrs	2.61	0.38	ns	1.72	1.60		
	> 3 hrs	2.82	-1.18	–	1.18	1.19		
Q8f				(0.0058)			ns	ns
Trust institutions: EU parliament?	0 No trust at all	4.86	4.11	0.017	7.49	0.07		
	1	4.60	-3.43	ns	2.28	0.02		
	2	7.96	-2.24	ns	6.41	0.04		
	3	8.74	-0.07	ns	8.65	0.04		
	4	9.98	-2.11	ns	8.39	0.02		
	5	19.06	6.63	ns	23.33	-0.01		
	6	12.79	-0.83	ns	12.37	-0.04		
	7	11.63	-2.49	ns	10.47	-0.05		
	8	11.88	3.07	ns	13.88	-0.07		
	9	4.99	-2.83	ns	3.07	-0.02		
	10 Complete trust	3.51	0.18	–	3.66	-0.02		
Q11				0.0000			0.000	0.000
Immigration: same ethnicity?	Allow many	19.81	8.36	0.004	19.41	9.38		
	Allow some	30.64	0.85	0.007	29.01	3.12		
	Allow a few	33.61	0.22	0.000	38.92	-7.73		
	Allow none	15.93	-9.41	–	12.66	-4.77		
Q12				0.0000			0.000	0.000
Immigration: different ethnicity?	Allow many	9.16	2.85	ns	7.86	5.22		
	Allow some	22.37	3.92	0.047	20.16	7.02		
	Allow a few	43.66	7.50	0.000	51.87	-4.71		
	Allow none	24.81	-14.28	–	20.11	-7.53		
Q14				0.0003			ns	ns
Immigration: impact on economy?	0 Bad for economy	10.86	1.53	ns	13.27	-2.36		
	1	7.89	-6.22	ns	4.06	-0.60		
	2	7.15	-3.14	0.012	5.56	-0.73		
	3	9.14	1.14	ns	10.49	-1.06		
	4	9.65	-0.88	ns	9.40	-0.58		
	5	28.96	3.34	ns	31.01	0.72		
	6	5.05	4.34	ns	7.40	0.86		
	7	7.27	0.73	ns	6.90	1.09		
	8	6.85	1.47	ns	7.15	1.47		
	9	2.22	-0.60	ns	1.53	0.36		
	10 Good for economy	4.96	-1.72	–	3.23	0.84		
				0.0000			ns	ns
Q17d Gender role: parents should not divorce?	Agree strongly	10.42	-4.42	0.017	7.29	0.29		
	Agree	12.01	-1.07	0.046	10.76	0.56		
	Neither	19.15	5.08	ns	22.12	1.09		
	Disagree	28.78	10.94	0.000	36.17	0.07		
	Disagree strongly	29.65	-10.54	–	23.66	-2.01		

[1] For each item the first row indicates the P-Value from a Wald test of the joint hypothesis that the mode coefficients for all dichotomisations are zero. P-Values in brackets were larger than 0.05 after adjusting for multiple comparisons using the Holm-Bonferroni procedure. Rows two and following indicate the P-Values of the mode coefficients for each dichotomisation.

Table 3: Effect of mode on relationship between summed immigration score and predictors

Predictor of	Telephone		Predictor Variable		Predictor*Telephone	
immigration score	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value
In work	0.044	0.013	-0.020	ns	0.020	ns
Voted	0.124	0.001	0.076	0.015	-0.081	0.036
Centre right party	0.060	0.003	0.001	ns	-0.033	ns
Internet	0.045	0.010	0.051	0.016	0.020	ns
Trust people	0.051	ns	0.018	0.000	0.000	ns
Life satisfaction	0.090	0.008	0.025	0.000	-0.006	ns
Religiosity	0.065	0.002	0.000	ns	-0.002	ns
Political interest	0.042	ns	-0.130	0.040	0.020	ns
Gender roles	0.042	ns	0.104	ns	0.015	ns

Notes: Coefficients from separate OLS models of summed attitude score on mode (face-to-face omitted), predictor variable, interaction of mode and predictor variable and socio-demographic variables.