Martin, Peter; Lynn, Peter

Working Paper

The effects of mixed mode survey designs on simple and complex analyses

ISER Working Paper Series, No. 2011-28

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

Suggested Citation: Martin, Peter; Lynn, Peter (2011) : The effects of mixed mode survey designs on simple and complex analyses, ISER Working Paper Series, No. 2011-28, University of Essex, Institute for Social and Economic Research (ISER), Colchester

This Version is available at:
https://hdl.handle.net/10419/65907

ISER Working Paper Series

The effects of mixed mode survey designs on simple and complex analyses

Peter Martin
Centre for Comparative Social Surveys
City University

Peter Lynn
Institute for Social and Economic Research
University of Essex

No. 2011-28
November 2011

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

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

www.iser.essex.ac.uk

Non-Technical Summary

Survey data can be collected in a number of ways. The survey organisation may use face-to-face interviewing or telephone interviewing, or may ask respondents to complete questionnaires themselves, either online or on paper. It is becoming increasingly common for surveys to use a combination of these methods, a so-called “mixed mode” design. However, the choice of mode, or mix of modes, can affect the data that are collected and consequently also the estimates that are made based on those data. This can happen because different kinds of people are more likely to participate in different modes, or because people will give slightly different answers depending on the mode of interviewing.

In this paper we compare estimates from surveys that used three different designs. One survey was a single mode survey, carried out entirely through face-to-face interviewing. The other two surveys were mixed mode surveys, involving face-to-face and telephone interviewing as well as online (web) questionnaires. The difference between the two mixed mode surveys was that in one case respondents were offered an explicit choice between the three interview modes, while in the other case respondents were first asked to complete a web questionnaire and only if they were unable or unwilling to do so were they then asked instead for a telephone interview. If the respondent was also unable or unwilling to carry out the telephone interview they were then asked for a face-to-face interview. Our main interest lies in comparing each of the mixed mode designs with the single mode design.

An identical questionnaire was administered to each of the three samples, namely the round 4 questionnaire of the European Social Survey. We compare the distributions of answers to each survey question, and we also compare some estimates of regression coefficients from statistical models of the kind often used by political scientists.

While many estimates show no significant difference between the survey designs, we conclude that some estimates are affected by the survey design. We find some suggestive evidence that this is likely to be, at least partly, due to differences in how people answer questions in different modes rather than due to different types of people taking part in the survey. We consequently urge caution in comparing estimates based on data collected using different (mixes of) modes and in the use of mixed mode survey designs.

The Effects of Mixed Mode Survey Designs on Simple and Complex Analyses

Peter Martin

Centre for Comparative Social Surveys, City University

Peter Lynn

Institute for Social and Economic Research, University of Essex

Abstract: We compare two alternative mixed mode survey designs with face-to-face data collection in terms of differences in estimates. Both mixed mode designs involve face-to-face, telephone and web interviewing. One design uses modes sequentially; the other offers respondents an explicit choice of mode. All three samples are probability samples of the Dutch population, selected from the same frame in the same way and administered the same survey instrument, namely the questionnaire of round 4 of the European Social Survey. We find differences and consequently urge caution in comparing estimates based on data collected using different (mixes of) modes.

Keywords: measurement error, mode effects, nonresponse bias, social desirability bias, telephone interviewing, web surveys

JEL Codes: C81, C83

Acknowledgments: This work was supported by funding from Sixth Framework Programme of the European Union for the European Social Survey Infrastructure project (Contract Number 026042). The data were collected as part of the European Social Survey Infrastructure Preparatory Phase project (Contract Number 212331). The work forms part of an ongoing European Social Survey programme of research into mixed mode data collection. We are indebted to Gillian Eva for managing the survey design and implementation and we are also grateful to GfK Netherlands for data collection, particularly Peter Willems and Peter van Eijk.

Contact: plynn@essex.ac.uk; peter.martin@annafreud.org

1. Introduction

Compared to a classic single-mode survey, a mixed-mode survey – that is, a survey where different respondents answer the same survey questions using different modes of data collection – brings with it a number of complications. Put succinctly, a mixed-mode survey means extra work for the questionnaire designer (who needs to design a separate instrument for each mode of data collection), the survey administrator (who needs to manage samples in different modes), and the data manager (who needs to develop algorithms to process information collected in different modes in a compatible way). The current report focuses on a fourth group: namely, the end user of the survey, the data analyst. The context of our inquiry is the European Social Survey, which has so far been run as a single-mode face-to-face survey, but which must soon face the decision of whether to allow other modes of data collection to be used alongside face-to-face interviews. The purpose of the European Social Survey (ESS) is to provide high-quality data on European social and political attitudes that can be analysed by researchers with a multitude of potential research questions in mind. Before making a decision about whether the ESS should allow some countries to collect data using mixed mode designs, rather than purely by face-to-face interviews, we need to consider the consequences such a change would have for the users of our data.

Within a given ESS country, switching from single-mode face-to-face data collection to mixed mode data collection may appear attractive for two kinds of reasons: first, mixed mode data collection has the potential to reduce data collection costs relative to a single-mode face-to-face survey; second, some authors have expressed the hope that, because individuals in a given population may differ in the likelihood to which they respond to different modes of data collection, mixed-mode data collection may help to increase survey response rates and reduce non-response error.

A mixed mode survey that indeed does turn out to have either or both of these advantages could have a benefit for data analysts. Reduction of data collection costs per respondent might allow a given country to access a larger sample of respondents; or it might indeed make possible the participation of a country that would otherwise be deterred by the high cost of face-to-face data collection. And a reduction of non-response error would benefit the validity of the data analyst's results.

Yet neither of these advantages is certain. Cost advantages of a mixed mode survey are difficult to calculate, as any reduction in fieldwork costs in a given country must be offset against increased costs of questionnaire development, survey administration, and data management. And while there is strong evidence that mixed-mode data collection can increase response rates compared to postal or telephone surveys, there is no evidence that mixed-mode surveys do better than single-mode face-to-face surveys in this respect (Dillman 2009, Lynn et al 2010, Martin 2011).

For the data analyst, then, the advantages of mixed mode data collection would seem to be tenuous at best. There is, however, a significant disadvantage associated with mixed mode data collection, and that is the threat it poses to equivalent measurement. Mixed mode data collection designs may result in different measurement error than face-to-face data collection in a variety of ways: through different population coverage, different mechanisms of respondent selection, and different measurement effects caused by the psychological implications of different data collection situations (such as the presence or absence of an interviewer, or the difference between aural and visual information processing).

Differential measurement error due to country differences in data collection design (for example, when some countries carry out the ESS as a face-to-face survey, and others as a mixed mode survey) is a complex issue to deal with for the data analyst. There are three main issue to consider:

1. Effects of data collection design are not uniform across variables. It will therefore hardly be feasible to give general advice to data analysts, who may be concerned about the internal validity of their research.

2. Based on current knowledge, mode effects are difficult to predict. Although it is true that some causal mechanisms of measurement effects are quite well understood in principle (for example, social desirability, primacy and recency effects), it is rarely clear how a given previously untested variable would be affected by data collection mode.

3. There is no way of 'adjusting' data for mode effects or effects of data collection design in a way that would be useful for all types of analysis. In particular, there is no adjustment method akin to the design weights that are used to adjust samples for the effects of unequal sampling probabilities. It may be possible, in some instances, to correct estimates of means or proportions for known mode differences in bias. However, this method will be complicated due to the issues summarized in points 1 and 2 above, and it will not address the problem of measurement differences that affect other types of estimates (such as correlations or coefficients in a statistical model).

On the other hand, while measurement differences between modes exist, it is not clear that they would be large from a substantive point of view. In particular, most observed large mode differences in measurement have been found to be reducible by careful questionnaire design (Dillman 2009).

How many ESS estimates would be affected by which types of measurement effects is, of course, an empirical question – as is the strength of these effects. While no single study will be sufficient to gauge the overall effect that the introduction of mixed mode data collection would have on ESS data quality, this report will attempt to gauge the seriousness of the loss of data quality, and in particular of the loss of cross-national and diachronic equivalence, by examining data from the ESSPrep Mixed Mode Experiment, conducted in the Netherlands alongside Round 4 of the ESS (Eva et al. 2010).

The aim of this study is to gauge the extent to which mixed mode data collection, if introduced in a single country, would compromise measurement equivalence in the ESS. We will focus on three types of measurements:

(1) Univariate distributions of all ESS variables;
(2) Attitude scales; and
(3) Multivariate analysis.

Unlike many studies that have investigated mode effects or the effects of mixed mode data collection, we do not limit our analysis to distributions of a small number of individual

variables, but aim to gauge the effect of mixed mode on a range of different types of variables, as well as their correlations, and the consequences of any observed effects for multivariate analysis.

2. Methods and Data

In this paper we are interested in assessing differences in estimates between data collected in the traditional ESS face-to-face mode and data collected under either of two mixed mode designs. One of the mixed mode designs involved a sequential use of telephone (CATI), web (CAWI) and face-to-face (CAPI) interviewing while the other offered respondents a concurrent choice of the same three modes. For brevity, we will refer to these designs as the sequential design and the concurrent design. The designs are described in more detail below. Our comparisons can be thought of, in the language of deLeeuw (2005), as survey protocol comparisons. We are interested in assessing the overall effect of using one survey protocol rather than another; we do not attempt to separate out the effects of particular modes or to separate effects on measurement from effects on selection. Our focus is on whether estimates are comparable when data are collected under the difference designs under consideration.

Data Collection Protocols

The face-to-face data for this study come from the Netherlands component of the fourth round of the European Social Survey (ESS), 2008-09. The mixed mode data come from an experimental study carried out in the Netherlands around the same time. The experimental study involved the selection of a random sample of the general population using the same sampling frame and same design as for the main ESS in the Netherlands. This sample was then randomly allocated to alternative mixed mode treatments and field work was carried out using the same survey instrument as the main ESS and by the same survey organisation.

The experimental sample was based on a one-stage systematic sample of 2,500 addresses selected from the 'Postaal Afgiftenpuntenbestand,' a list of locations to which the Dutch mail service delivers mail. The next step was to attempt to match phone numbers to the sample addresses. This was done using an automated matching service that the survey agency, GfK,

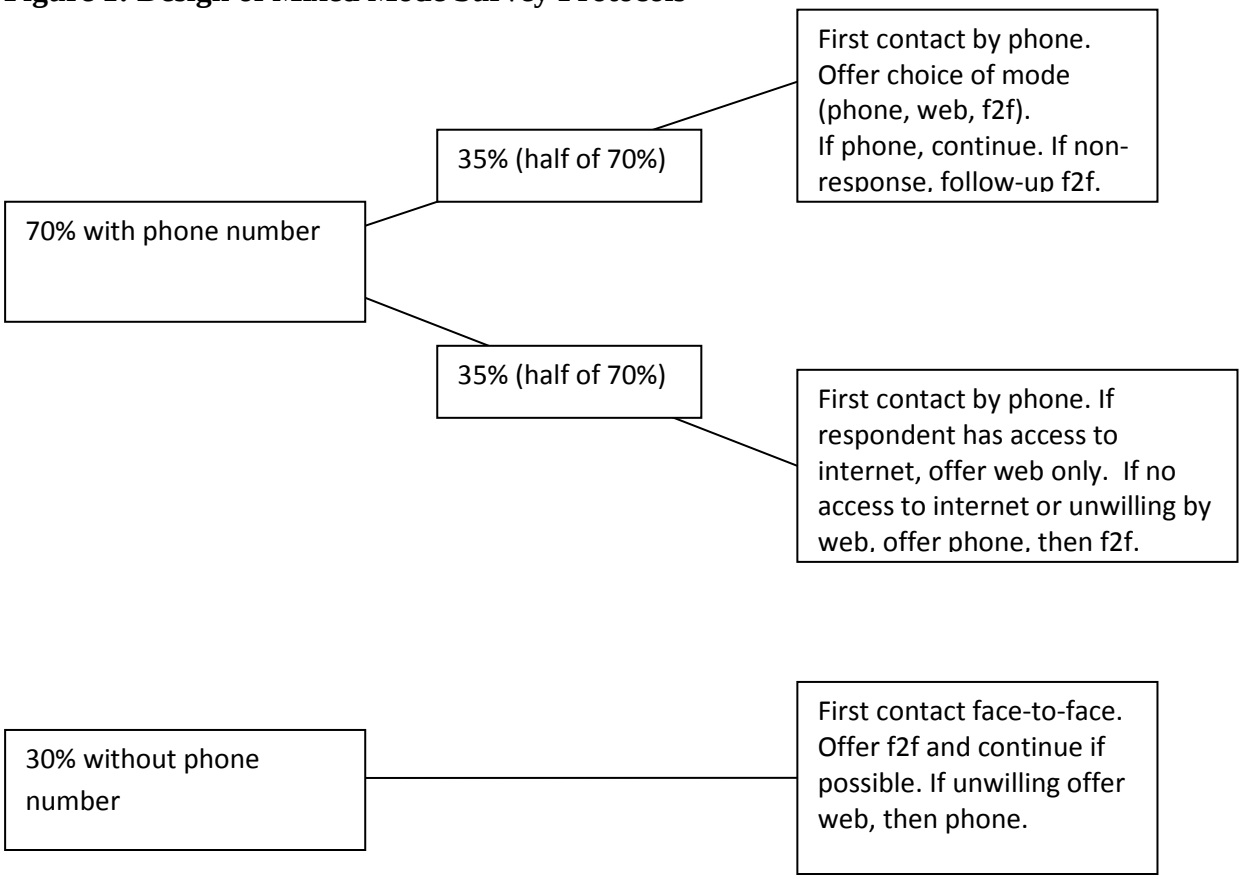
uses regularly to enhance address lists and which operates via an intermediate step of first matching names to addresses. However, it was anticipated that only around 70% of sample addresses would be successfully matched to a phone number by this process. In the event, the match rate was 70.2%. The successfully matched addresses were then randomly split into two groups, one to be administered the sequential mixed mode design and the other to be administered the concurrent mixed mode design. In both cases the modes were to be CATI, CAWI and CAPI. For the sequential group, CAWI was to be attempted first, followed by CATI and finally by CAPI. For both groups all initial contacts would be attempted by telephone. Though this was expected to be less successful for CAWI participation than an email approach (Messer & Dillman, 2009), email was not an option as email addresses were not available. The approach used was considered to be the best that was feasible. The 30% of sample addresses for whom telephone numbers were not matched were all treated in the same way, namely a sequential mixed-mode design that involved first offering CAPI during a face-to-face approach, then CAWI (during the same approach, if CAPI was refused). Finally, CATI was offered as a last resort, though this involved collecting a phone number on the doorstep and was not expected to be a popular option with sample members. This design is summarised in Figure 1.

Fieldwork for the experimental sample was carried out between 24 November 2008 and 7 July 2009, while fieldwork for the main ESS sample took place between 7 September 2008 and 27 June 2009. It is important to note that the ESS interview lasts around one hour on average, so this is a long instrument in the context of CATI and CAWI. In consequence, our study provides a rare comparison of mixed mode designs for a typical face-to-face survey instrument. This reflects the reality of the situation faced by the ESS. Most other mixed mode studies that involve at least one self-completion mode have been limited to much shorter instruments.

The main ESS survey achieved a response rate of 52.0% The experimental mixed mode survey achieved response rates of 44.8% for the sequential mixed mode design (n=363 respondents), 45.9% for the concurrent mixed mode design (n=367 respondents), and 39.5% (n=267 respondents) amongst the sample for whom telephone numbers could not be matched. For analysis purposes, respondents for whom telephone numbers could not be matched were randomly allocated to either the sequential or concurrent samples, thus simulating the

samples that would be achieved with either mixed mode design in the real situation where 30% would not be successfully matched to a telephone number. With the samples combined in this way, the overall response rates were 43.4% for the sequential design and 44.0% for the concurrent design. Of all interviews achieved, in the sequential design 39% were CAWI, 18% CATI and 44% CAPI. With the concurrent design, 36% were CAWI, 18% CATI and 45% CAPI.

Figure 1: Design of Mixed Mode Survey Protocols



Even though the response rates differed between the face-to-face survey and the mixed mode designs, an analysis of demographic composition of the three samples found no significant differences with regard to the variables age, sex, education, household size, and degree of urbanisation in the area of residence (Eva et al. 2010; Vannieuwenhuyze et al. 2010), although the demographic composition of all three samples differed from Dutch population statistics (Eva et al. 2010). Thus, although none of the three samples are perfectly representative of the population they were drawn from, we find no evidence of an effect of

survey protocol on the demographic composition. This finding does not, of course, rule out the possibility that the three different survey protocols may have systematically attracted different types of respondents with respect to characteristics other than age, sex, education, household size, and degree of urbanisation.

Analysis Methods

We carry out three sets of analyses. The first (section 3 below) involves comparing sample means or distributions between the single-mode (face-to-face) design and each of the mixed mode designs. We do this for 230 items[1] and for each of the two mixed mode designs (sequential and concurrent), so 460 comparisons in total. For each comparison we use a regression modelling approach to test whether design has a significant effect on the mean or distribution of the item. The details of how we do this are set out in section 3.

The second set of analyses (section 4 below) recognises the fact that surveys frequently aim to measure concepts through more than one variable. In particular, attitudes are often measured by three or more indicators that load on a common underlying, latent factor. In statistical analysis, it is often the factor score – rather than either individual variables or a straightforward summation of variable scores – that is used as a measure of the attitude. We therefore investigate whether measurement of latent variables is equivalent across data collection designs.

We consider six attitude scales plus the Schwartz Human Values scale. To assess whether these seven concepts are measured equivalently across data collection designs, we perform multi-group confirmatory factor analysis, treating each data collection design as one group, and tested for configural, metric, and scalar invariance across groups. When all three types of invariance are found, we can conclude that concepts are measured equivalently across

[1] The questionnaire consisted of 235 items. We excluded 5 of them from our analysis because their distributions meant that any statistical comparison between samples would be extremely unreliable. This was the case for the variable "Party Membership", which is a nominal variable with many categories where statistical models run into the problem of low cell counts; and for four variables that count "Number of employees" and "Number of people responsible for at job" each for the respondent and the respondent's partner (if such a partner exists). These four questions are count variables with extremely skewed (non-normal) distributions that are heavily affected by individual outliers. For these last four variables, we considered non-parametric tests, but these are not supported by STATA's "svy" commands, which we used to apply design weights and adjust standard errors of estimates accordingly.

designs. This method of testing measurement equivalence has become standard in research on the performance of measurement instruments across cultural groups (van de Vijver 2011). The seven scales and details of the analysis methods are described in section 4 below.

ESS data users are rarely interested in estimating univariate distributions, however. Users are primarily interested in understanding the relationships between variables and how these relationships differ between population subgroups. Consequently, the data are typically used to estimate multivariate models of various kinds. So, our third set of analyses test whether estimated multivariate models differ when using data collected by mixed mode protocol rather than face-to-face interviewing. We cannot claim that this testing is in any way comprehensive or representative of models in which ESS users may be interested, but we hope it is illustrative of possible effects, The details of the modelling and testing approaches used are set out in section 5 below.

3. Univariate Analysis

In the first stage of our analysis, we examine the univariate distributions of all items contained in the ESS Round 4 questionnaire. For each item we fit an appropriate regression model, with the survey item as the dependent variable and a dummy variable "survey design" as the sole independent variable, where "survey design" is coded "1" for cases from the mixed mode sample, and "0" for cases from the face-to-face survey. This we do twice for each item: first, comparing the sequential mixed mode design with the face-to-face survey, and second, comparing the concurrent mixed mode design with the face-to-face survey. The type of regression model differs with the measurement level of the item under investigation: for items with 6-point-scales, 11-point scales or ratio-scale measurement, we use ordinary least squares regression; for items with 4- or 5-point scales and other ordinal scales we use ordinal logit regression; for dichotomous items we use logistic regression; and for nominal items multinomial regression. The criterion for inferring an effect of survey design was given by the F-test comparing the model with a null model without the dummy predictor. In the case of OLS regression, this test is equivalent to a t-test for independent samples; in the case of ordinal and binary logistic regression, the test is equivalent to a z-test on the coefficient for our dummy "survey design". The data were weighted, using weights inversely proportional to

selection probability, and the “svy setup” command in STATA was used to apply the appropriate weights.

As we compare 230 variables between survey designs, with an α-level of .05, we would expect 11.5 results to appear significant for each mixed mode design simply by chance, even if there were no true differences. As Table 1 shows, however, the number of significant results we actually observed was considerably larger in both cases. Compared to the face-to-face sample, the Concurrent Mixed Mode Sample featured differing distributions in 25 items, while in the Sequential Mixed Mode Sample 38 items had distributions significantly different from the face-to-face sample.

The experiment allows us only to examine the overall effect of switching from the single mode face-to-face design to a mixed mode design. The effects that we did find might have resulted from one or both of the following types of causes: (1) Selection effects at the survey response level, and (2) measurement effects associated with the differences between the face-to-face mode and the other modes used in the experiment. Since the mixed mode design included the use of three different modes of data collection, any net measurement effect for the sample as a whole may be a combination of the difference between face-to-face and telephone modes on the one hand, and between face-to-face and web on the other.

Since selection and measurement effects are confounded in our experiment, we cannot strictly say anything about the causes of the differences we found. However, we will examine our results with a view to detecting potential patterns that may serve as hypotheses to be tested against evidence from other published studies, or by future research.

Table 1: Variable-by-variable comparison of each mixed mode design with the single-mode design

Type of Variable	Statistical Model	Number of variables	Number of significant sample differences: Concurrent mixed mode versus face-to-face	Number of significant sample differences: Sequential mixed mode versus face-to-face	Expected differences (per comparison)
Open numerical	Linear (t-test)	7	1	1	0.35
Attitude - 11-point	Linear (t-test)	60	3	10	3
Values - 6 pt	Linear (t-test)	21	5	3	1.05
Attitude - 4-5 pt scale	Ordinal logit	63	10	13	3.15
Behavioural - ordinal	Ordinal logit	13	2	3	0.65
Knowledge etc - ordinal	Ordinal logit	10	1	3	0.5
Status & relationships - ordinal	Ordinal logit	12	1	0	0.6
Behaviour - Nominal	Multinomial logit	2	0	0	0.1
Behaviour: binary	Logistic regression	10	0	1	0.5
Status & relationships: binary	Logistic regression	32	2	4	1.6
Total	**-**	**230**	**25**	**38**	**11.5**
Probability of result (assuming null hypothesis: no effect of survey design)			**<.001**	**<.001**	**--**

One consideration makes certain measurement effects plausible: since both mixed mode samples have a higher proportion of web respondents than telephone respondents, some measurement effects might be influenced more strongly by web effects than telephone effects. For example, many studies have shown that social desirability bias is weakest in self-completion modes (such as web questionnaires), stronger in face-to-face interviews, and strongest in telephone interviews (Kreuter, Presser, and Tourangeau, 2008; Tourangeau and

Yan, 2007; but see also Bowling, 2005). If some ESS variables are affected by social desirability bias, we would expect that relative to face-to-face interviews (the mode with which we would ideally like to maintain measurement equivalence), the web mode would decrease the bias, whereas the telephone mode would increase it. Overall, because of the greater proportion of web respondents compared to telephone respondents, we would expect the mixed mode samples to have slightly lower social desirability bias than the face-to-face sample.

We identified 19 items which appear to have social desirability connotations. They are listed in Table 2. Of these, ten items exhibit significant differences between one or both mixed mode sample and the face-to-face sample. This is a far greater proportion of significant results than we would expect to find by chance alone. We suspect that these differences may have been caused by social desirability effects. These questions cover diverse topics: television consumption, interest in politics, attitudes to migration, social contact, and age prejudice. In all ten of these variables, the sequential mixed mode sample differed significantly from the face-to-face sample, as mixed mode respondents were more likely to report socially undesirable attitudes and behaviour than single-mode face-to-face respondents. The concurrent mixed mode sample differed from the sequential sample in the same direction on all variables, but the difference reached significance in only one case (television consumption). While these results seem to confirm the hypothesis of reduced social desirability bias in the mixed mode estimates compared to the face-to-face estimates, it is puzzling that the mixed mode samples should differ in the strengths of these effects, since the distribution of modes in the two samples was nearly identical.

Table 2: Questionnaire items with suspected social desirability connotations

Item	Question Wording	Direction of difference	Statistically significant result in sample
A1	On an average weekday, how much time, in total, do you spend watching television?	MM: watch more TV	Sequential & concurrent
B1	How interested would you say you are in politics …?	MM: less interest in politics	Sequential
B35	To what extent do you think [country] should allow people of the same race or ethnic group as most [country]'s people to come and live here?	MM: allow fewer immigrants	Sequential
B36	How about people of a different race or ethnic group from most [country] people?	MM: allow fewer immigrants	Sequential
B37	How about people from the poorer countries outside Europe?	MM: allow fewer immigrants	Sequential
B38	Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?	MM: immigration worse for economy	Sequential
B39	Would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?	MM: immigration rather undermines cultural life	Sequential
B40	Is [country] made a worse or a better place to live by people coming to live here from other countries?	MM: immigration makes country worse place	Sequential
C3	Do you have anyone with whom you can discuss intimate and personal matters?	MM: less likely to have anyone	Sequential
E53	Please tell me how important it is for you to be unprejudiced against people of other age groups.	MM: consider it less important to be unprejudiced	Sequential
B2	How often does politics seem so complicated that you can't really understand what is going on?	(no effect)	none
B3	How difficult or easy do you find it to make your mind up6 about political issues?	(no effect)	none
B11	Some people don't vote nowadays for one reason or another. Did you vote in the last Dutch national election [...]?	(no effect)	none
B24	All things considered, how satisfied are you with your life as a whole nowadays?	(no effect)	none
B31	Gay men and lesbians should be free to live their own life as they wish1.	(no effect)	none
C1	Taking all things together, how happy would you say you are?	(no effect)	none
D3	A woman should be prepared to cut down on her paid work for the sake of her family.	(no effect)	none
D6	When jobs are scarce, men should have more right to a job than women.	(no effect)	none
F32	[What is your] household's total income, after tax and compulsory reductions, from all sources?	(no effect)	none

Note: MM: Mixed mode sample. The full questionnaire can be viewed and downloaded at www.europeansocialsurvey.org .

Table 3: Sample differences in items with pessimism connotations

Item	Question Wording	Direction of difference	Statistically significant result in sample
A8	Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?	MM: Report less Trust	**Sequential**
A9	Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?	MM: Report less Trust	**Sequential**
C7	How often, if at all, do you worry about your home being burgled?	MM: report more worry about burglary	**Concurrent and Sequential**
C9	How often, if at all, do you worry about becoming a victim of violent crime?	MM: report more worry about violent crime	**Sequential**
C15	How is your health in general?	MM report worse health	**Concurrent**
D14	And what do you think overall about the opportunities for young people to find their first full-time job in [country]?	MM judge opportunities to be worse	**Concurrent and Sequential**
D43	There are insufficient benefits in [country] to help the people who are in real need.	MM: more likely to agree that benefits are insufficient	**Sequential**
D47	Please tell me how likely it is that during the next 12 months you will be unemployed and looking for work for at least four consecutive weeks?	MM: believe unemployment to be more likely	**Concurrent and Sequential**
D49	And during the next 12 months how likely is it that there will be some periods when you don't have enough money to cover your household necessities?	MM: believe themselves more likely to be in financial difficulties	**Concurrent and Sequential**

Another pattern that emerged from the inspection of results was a tendency of mixed mode respondents to give more pessimistic answers to a variety of questions that invited respondents either to evaluate the present, or to gauge risks and opportunities. Table 3 shows nine variables that correspond to this pattern. We found no variable where the mixed mode

sample exhibited the opposite tendency (toward optimism). The cause of this difference must remain conjecture, but there is some evidence that self-completion modes engender more pessimistic answers on subjective measures of health than face-to-face interviews (Christensen, Ekholm, and Juel 2011). It is possible, then, just as we speculated in the case of the variables with social desirability connotations, that those respondents who chose the web mode may have 'pulled' the means of the mixed mode samples in the direction of more pessimistic answers.

4. Investigating Scale Equivalence

In section 3 we compared the univariate distributions of individual survey items measured through mixed mode data collection and face-to-face data collection. In this section we investigate whether measurement of latent variables is equivalent across data collection designs. Each latent variable is measured by a factor score from a scale of items. The seven scales that we investigate comprise six attitude scales, plus the Schwartz Human Values scale, which consists of 21 questions designed to measure ten basic human values (which in turn are organized into four higher-order values). The selected scales are presented in Table 4.

Table 4: Attitude Scales for which measurement equivalence is tested

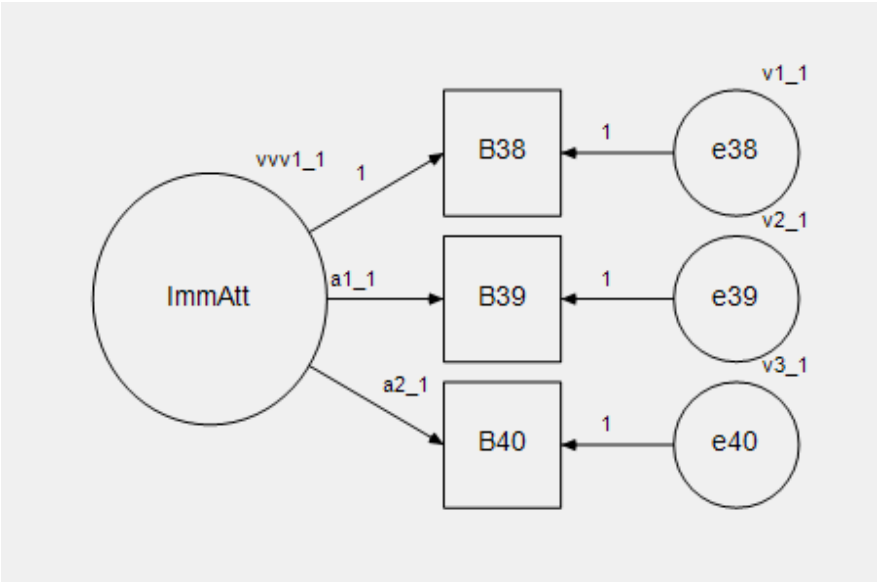
Scale	Items	Reference
Social Trust	A8 – A10	Allum, Read & Sturgis (2011)
Political Trust	B4, B7, B8	Allum, Read & Sturgis (2011)
Political Efficacy	B1 – B3	Halperin & Sulitzeanu-Kenan (2010)
Attitude to Immigrants	B38 – B40	Martin (2010)
Attitude to Immigration	B35-B37	Meuleman & Billiet (2008)
Religious Involvement	C21 – C23	Meuleman & Billiet (2011)
Human Values	G_A – G_U	Schwartz (2007)

To investigate whether these seven concepts are measured equivalently across data collection designs, we perform multi-group confirmatory factor analysis, treating each data collection design as one group, and test for configural, metric, and scalar invariance across groups.

Configural invariance refers to the assumption that items measure the same concepts across groups. If this assumption holds, items should load on the same factors across groups. We test this assumption by performing, separately for each group, an exploratory factor analysis on the items of each factor. If all items load on the same factors across groups, we conclude that there is configural invariance between designs.

Metric invariance refers to the assumption that the scale measuring the concept has the same measurement units across groups. This means that even if a certain source of bias (say, social desirability bias) affects one group more than the other, the relative scores of individuals within each group are not affected (cf. Van den Vijver 2011, p. 9). Metric invariance assumes configural invariance. Operationally, we define metric invariance as achieved when the factor loadings of all items in a model do not differ statistically across groups. To see what this means, consider Figure 2.

Figure 2: Confirmatory factor analysis with three observed variables and one latent factor



The factor loadings are symbolized by the three arrows going from the latent factor (here "ImmAtt") to the three observed variables B38, B39, and B40. To identify the model, one factor loading has to be set to the value 1. The other two, denoted here by "a1_1" and "a2_1" are estimated separately for each group. If metric invariance holds, we can constrain the two

loadings “a1_1” and “a2_1” to be equal in all groups (i.e., all data collection designs) without loss of model fit relative to a model where all factor loadings are free to vary across groups.

Scalar invariance refers to the assumption that the concept is measured on the same interval or ratio scale in all groups, so that a given score on the scale has the same meaning across groups. It is necessary to assume scalar invariance if we want to make meaningful comparisons across groups. Thus, scalar invariance is necessary if we want to claim equivalence of measurement across data collection designs. Scalar invariance presupposes both configural and metric invariance. Operationally, we define scalar invariance as achieved if the intercepts of the observed variables in the confirmatory factor analysis illustrated in Figure A can be held constant across groups without compromising the model fit.

To test for metric and scalar invariance, we proceed as follows: for each scale, we fit a multigroup confirmatory factor analysis model with three groups: ‘concurrent mixed mode’, ‘sequential mixed mode’, and ‘single mode face-to-face’. First we estimate an unconstrained model, leaving all parameter estimates free to vary across group. In a second step, we estimate a “metric invariance” model, fixing the unstandardized factor loadings to be equal across groups. Finally, we estimate a “scalar invariance” model by fixing the unstandardized factor loadings and the intercepts of the observed variables to be equal across groups. These three models are nested (i.e., each subsequent model is a special case of the previous model). Therefore, we can compare any two of our models using a chi-square test of the difference of the model chi-square values, which were each computed from the differences between the observed covariances and the covariances implied by the model. We use an alpha-level of $\alpha \leq .01$ to determine whether there is a significant difference between the more restrictive and less restrictive model in the extent to which the observed covariances are reproduced.

Table 5 shows the results for the seven scales. We see that five of the seven scales display scalar invariance across data collection designs: Social Trust, Political Trust, Political Efficacy, Religious Involvement, and Attitudes to Immigration. Looking at the scale Attitude to Immigrants, however, we find that, by the chi-squared criterion, only configural and metric invariance are given, but not scalar invariance. Finally, for the Human Values Scale, only configural invariance is given, but neither metric nor scalar invariance.

Table 5: Results of chi-square tests of nested model comparisons

Scale	Number of Variables	Measurement invariance across designs?		
		Configural	**Metric**	**Scalar**
Social trust	3	n.s.	n.s.	n.s.
Political trust	3	n.s.	n.s.	n.s.
Political efficacy	3	n.s.	n.s.	n.s.
Religious involvement	3	n.s.	n.s.	n.s.
Attitude to Immigration	3	n.s.	n.s.	n.s.
Attitude to Immigrants	3	n.s.	n.s. (RMSEA<.001)	**p=.008** (RMSEA=.019)
Human values	21 (4 factors)	n.s. (RMSEA=.040)	**p=.007** (RMSEA .039)	**p<.001** (RMSEA=.038)

Note: Model comparisons: the "configural Invariance model" is tested against the "independence model" (assuming no correlations among the observed variables); the "metric invariance model" is tested against the "configural invariance model"; and the "scalar invariance model" against the "metric invariance model".
Human Values: We fitted a four-factor model, as predicted by Schwartz' theory of human values, using the higher-order factors Self-Enhancement, Conservation, Openness to Change, and Self-Transcendence.
Religious Involvement: As two of the three observed variables are measured on ordinal measurement scales, we used a weighted least squares (WLS) estimation procedure which analyses polychoric correlations and asymptotic covariance matrices, rather than regular covariance matrices. This replicates the model used by Meuleman & Billiet (2011, p. 187).

However, we can doubt whether it is wise to rely only on the chi-square test of model difference as an indicator of model fit. One of the disadvantages of this test is its sensitivity to sample size, so that with large sample sizes, even very small differences between models may result in significant chi-squared values. In the cases where a significant difference between models was found, we therefore consulted the RMSEA fit index as a check on our results. The RMSEA (Root Mean Square Error of Approximation) is a function of the chi-square value of the model, the sample size and the degrees of freedom of the model comparison. Simply put, the RMSEA offsets the chi-square value against the sample size and the degrees of freedom. By convention, many researchers use the criterion that if RMSEA<.05, the model fit is considered acceptable. However, simulation studies do not support the use of any single cut-off point for the RMSEA statistic (Chen et al. 2008). As it turns out, all our "scalar

equivalence" models have RMSEA<.05, including the models that, by the chi-square criterion, did not appear to fit the data. In the case of the Human Values Scale, a comparison of RMSEA values across the three models reveals that the "scalar equivalence" model has the lowest (that is, best) RMSEA value. In the case of the Attitudes to Immigrants Scale, the "scalar equivalence" model has the highest RMSEA of all three models (namely, RMSEA=.019 in this case). So for the Attitudes to Immigrants Scale, more than for the other scales, we may doubt whether scalar equivalence holds, although in practice most researchers, using current conventions, would probably regard the scalar equivalence model as having acceptable fit.

On balance, the results of this analysis are encouraging for those interested in using mixed mode data collection in the ESS. In all of the seven scales investigated, scalar invariance appears to be given, or approximated closely enough, so that latent variables obtained from mixed mode samples may be compared with the corresponding latent variables from face-to-face samples. Put differently, it appears that the correlational structure of variables designed to measure latent concepts would be little disturbed by a switch from face-to-face to mixed mode data collection.

5. Multivariate Analysis

In section 3 and 4 we assessed the impact of mixed-mode designs, relative to the single-mode face-to-face design, on univariate distributions for individual survey items and on factor scores. However, while this provides a useful initial insight into possible impacts on the data, few if any users of ESS data are interested in estimating univariate distributions. Commonly, the data are used to estimate multivariate models of various kinds. It is therefore desirable to understand the impact of mixed-mode designs on the estimation of such models. An impact on marginal distributions does not necessarily imply an impact on multivariate structure and, indeed, some studies have found significant effects of survey modes on univariate distributions but not on estimates of association between variables (e.g. Jäckle et al, 2006).

It is not feasible to attempt a comprehensive evaluation of all types of models that may be fitted to the ESS data. Instead, we focus here on the use of one particular variable that

exhibited significant differences in the univariate comparisons, namely interest in politics. The ESS measure of interest in politics is an important and widely-used measure, featuring both as an object of analytical interest in its own right and as an explanatory variable in the study of a range of phenomena. Furthermore, the differences in the distribution of this variable between the mixed mode and face-to-face protocols is likely to have a measurement component (Roberts et al 2006; Vannieuwenhuyze et al. 2010) associated with greater social desirability bias in face-to-face interviewing. Observed differences are in the direction consistent with this hypothesis (greater political interest expressed in face-to-face interviews). This could therefore be a test case for the impact of social desirability related mode effects on multivariate analysis.

We therefore explore the use of political interest both as a dependent variable (section 5.1) and as an independent variable (section 5.2) in substantive analytical models. In both cases our interest lies in whether the fitted model would be different depending on which data collection protocol had been used. As univariate comparisons suggested that data from the two mixed mode protocols were similar – perhaps unsurprising considering the similar distribution of modes and similar response rates under the two protocols – we have combined the data from both mixed mode protocols. We therefore compare mixed mode with face-to-face.

5.1 Determinants of Political Interest

Following Gabriel & van Deth (1998), we attempt to identify predictors of political interest. Gabriel & van Deth investigated religiosity, political libertarianism, left-right materialism, new egalitarianism, and materialism vs. postmaterialism. Additionally, they included sex, age and education as controls and refer to these as standard controls in the analysis of political interest (Gabriel & van Deth 1998, p. 399). The ESS4 questionnaire does not include all the variables used by Gabriel and van Deth, but for three of their five value dimensions we have indicators (or proxies). We also test some additional variables that to some degree might be able to replace the missing conceptual indicators, as well as including age, sex and education as controls. The variables we used are documented in annex A.

In two steps we assess the effect of data collection protocol on ordinal logit models of political interest. The first step involves fitting a series of models, each of which contains one or more predictor variables along with their interactions with 'sample', a binary indicator of whether the respondent is a member of the face-to-face sample or the mixed mode sample. Significant interaction terms indicate that a model of the relationship between political interest and the predictor variable(s) would differ depending on which data collection protocol had been used. Results are summarised in Table 6. At the second step, all predictor variables with a significant main effect or interaction, plus the demographic control variables, are included in a single model of political interest.

We find only two significant interaction effects. These both related to indicators of "Left-right Materialism", namely the variables D1 and D4. The main effects of D1 and D4 on political interest are not significant in either sample. However, these effects take opposite directions, leading to the result that when we combine the samples and examine the interaction between "sample" on the one hand, and "D1" and "D4" respectively, on the other, these interaction effects turn out to be statistically significant. The analysis of both variables indicates that "economic egalitarianism" is positively associated with political interest in the MM sample, but negatively in the F2F sample. This is true whether or not we control for age, sex, and education.

We found no evidence that the relationships of other variables to political interest were affected by data collection design. The demographic variables age, sex, and education are significant predictors of political interest, but their coefficients do not differ significantly between the two samples.

The effects of the other predictors can be summarised as follows. Approval of income redistribution to reduce inequality (B30) was negatively associated with political interest in both samples (albeit this association was stronger in the F2F sample than in the MM sample). Religiosity was not related to political interest in either sample. Social liberalism (represented by the proxy 'acceptance of homosexuality') was positively related to political interest in both samples, with no statistically significant difference in coefficients. Political liberalism (represented by the proxy 'should antidemocratic parties be banned') was not related to political interest in either sample. Gender traditionalism (represented by the variables 'men

should have more right to jobs' and 'women should cut down on paid work for sake of family') was negatively associated with political interest in both samples, with no significant difference in coefficients. Subjective left-right placement had no significant effect on political interest in either sample. Authoritarianism (represented by the variable "children in schools should be taught authority") did not have an effect on political interest in either sample. Law-and-Order orientation (represented by 'Harsher sentences') had a negative association with political interest, but no significant difference in coefficients.

A final model in which all significant predictors and interactions (when tested separately using Wald tests) are included is shown in Table 7. The nature of the two interactions with sample in this model, as described earlier, is illustrated in Figure 3 and Figure 4.

Table 6: Ordered Logit Models Predicting Political Interest

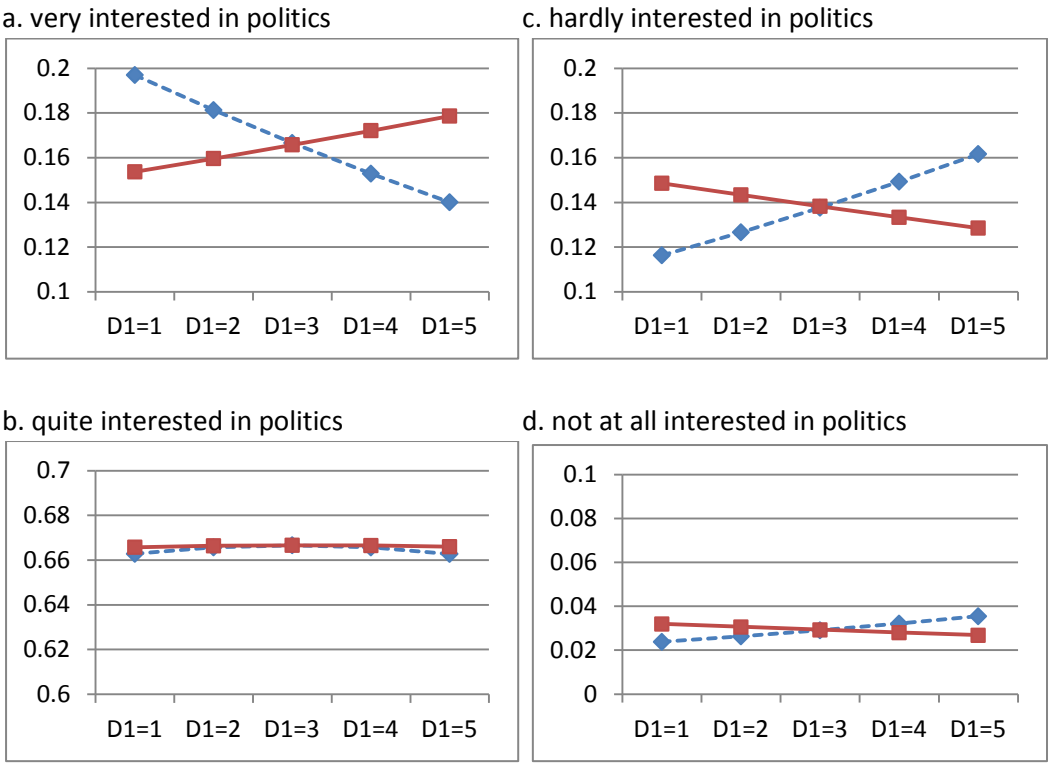
		Adjusted Wald-test		
Concepts	**Interaction**	**(df)**	**F**	**p**
Sex	Female*MM	(1, 2772)	0.19	.665
Age	(Age+Age^2)*MM	(2, 2769)	1.46	.233
Sex, Age	(Female+Age+Age^2)*MM	(3, 2768)	1.11	.345
Education	edulvl*MM	(2, 2762)	.68	.509
Sex, Age, Education	(Female+Age+Age^2+edulvl)*MM	(5, 2757)	.88	.491
Religiosity	C21*MM	(1, 2761)	.63	.428
	C22*MM	(5, 2768)	1.10	.359
	C23*MM	(5, 2762)	.98	.429
Left-right materialism	B30*MM	(1, 2749)	1.01	.315
	D1*MM	(1, 2746)	5.23	.022
	D4*MM	(1, 2747)	5.78	.016
New egalitarianism	B31*MM	(1, 2761)	.97	.324
	D3*MM	(1, 2758)	.04	.838
	D6*MM	(1, 2753)	.81	.367
Other	B32*MM	(1, 2720)	.45	.501
	D2*MM	(1, 2768)	.02	.900
	D5*MM	(1, 2745)	.78	.377
	B23*MM	(1, 2667)	.70	.403
Left-right materialism, demographic controls: age, age^2, sex, edulvl	(D1+D4)*MM	(2, 2716)	5.23	.005
Left-right materialism, demographic & attitude controls: age, age^2, sex, edulvl, B30, B31, D3, D5, D6	(D1+D4)*MM	(2, 2652)	4.43	.012

Notes: Each row corresponds to a separate model. The Adjusted Wald Test tests the hypothesis that all interaction parameters in the model are zero. Age and Age^2 were centred at their respective means to facilitate coefficient interpretation. Replacing Age and Age^2 with a categorical variable indicating age groups yields substantially the same results as shown here. The variable Edulvl is a simplified recode of F6 (Hqual). Using F6 instead (which has 14 categories) leads to substantially the same results. "MM" refers to a dummy variable that is coded 1 for the mixed mode sample, and 0 for the face-to-face sample. Please refer to Annex A for explanations of variable abbreviations

Table 7: Ordered Logit Model of Political Interest

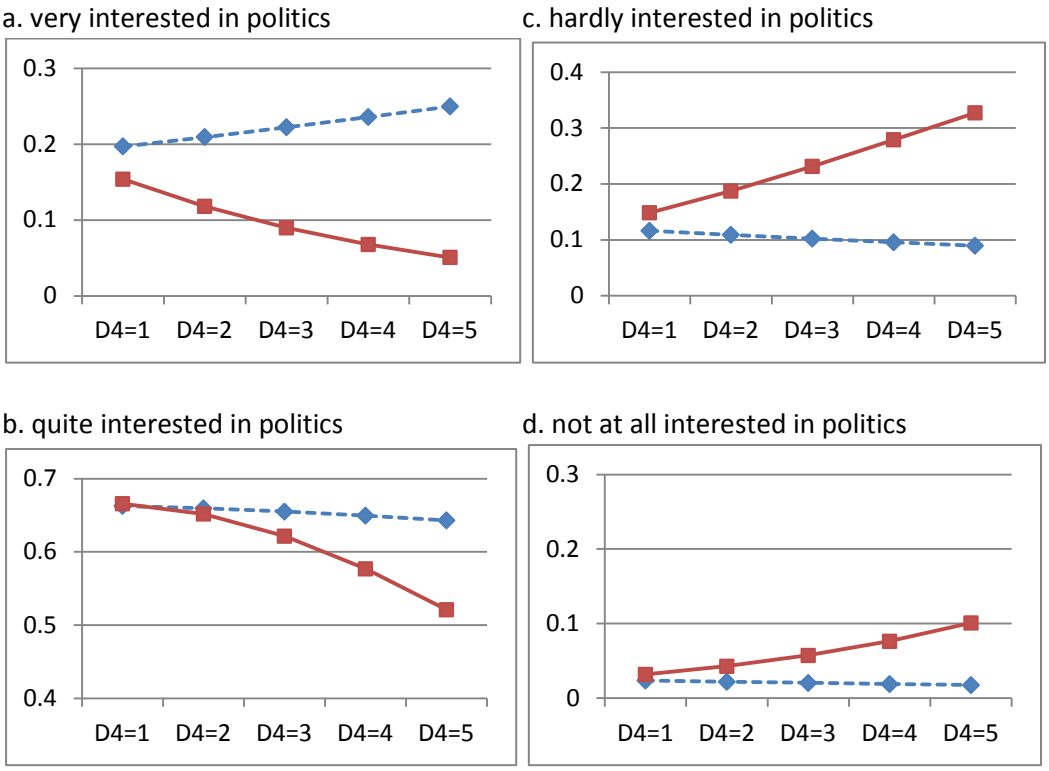
Variable	Coefficient	*P*
Sample (mixed mode)	0.0659	0.88
Sex (female)	0.5877	<0.001
Age (relative to mean)	-0.0223	<0.001
Age squared (relative to mean)	0.0002	0.13
Education: medium	-0.8555	<0.001
Education: low	-0.9065	<0.001
Left-right materialism 1 (B30)	-0.0808	0.10
New egalitarianism 1 (B31)	-0.0085	0.90
New egalitarianism 2 (D3)	-0.0597	0.18
New egalitarianism 3 (D6)	-0.1780	0.001
Harsher sentences (D5)	-0.2372	<0.001
Left-right materialism 2 (D1)	0.1023	0.11
D1*Mixed mode	-0.1474	0.14
Left-right materialism 3 (D4)	-0.0381	0.57
D4*Mixed mode	0.1911	0.05
Cut 1	-3.9252	<0.001
Cut 2	-0.7064	0.08
Cut 3	1.1903	0.003

Figure 3: Predicted Probabilities of Political Interest: Interaction between D1 and Mode



Notes: Y-axis plots the predicted probability of the given level of political interest; D1=5 indicates the most materialistic response; Dashed line indicates face-to-face, solid line indicates mixed mode.

Figure 4: Predicted Probabilities of Political Interest: Interaction between D4 and Mode



Notes: Y-axis plots the predicted probability of the given level of political interest; D4=1 indicates the most materialistic response; Dashed line indicates face-to-face, solid line indicates mixed mode.

5.2 Political Interest as a Determinant of Voting, Voluntary Activity and Political Involvement

Several authors have found that expressed interest in politics is a powerful predictor of propensity to turn out to vote in elections (e.g. Butler and Stokes 1971; Clarke et al 2003). Others have suggested that interest in politics should predict involvement in voluntary activity and in other forms of political participation.

In this section we develop a model to explain each of four forms of behaviour by interest in politics and a set of demographic characteristics. The four dependent variables are self-reported turnout, involvement in voluntary activity, broad political participation and active political participation. These four variables are defined below. For each model, we then test whether the predicted coefficients depend upon the data collection protocol by fitting interaction terms with a binary indicator of whether the respondent was a member of the uni-

mode face-to-face (main ESS) sample or the mixed-mode experimental sample. We discuss each of the four models in turn.

Vote Turnout

Our dependent variable is a binary indicator of whether or not the respondent reported having voted in the last national election (item B11). Respondents who reported being ineligible to vote (4.5%) are excluded from the analysis. Although self-reported turnout tends to suffer from over-reporting due to social desirability bias (Swaddle and Heath, 1989; Voogt and Saris, 2005), Clarke et al (2003) found that the predictors of turnout differed little between a model in which the unadjusted self-report measure was used as the dependent variable and one in which it was adjusted for likelihood of turnout. We therefore predict unadjusted self-report based on political interest, gender, age (in seven categories), level of education (3 categories), marital status and economic activity status. We select predictor variables using a forward stepwise procedure and an inclusion criterion of $P < 0.05$. The final model includes gender, age, education, marital status and activity status. When we then test for interactions with sample, we find only a significant interaction with age: the predicted probability of voting is significantly lower in the mixed-mode sample for 45-74 year-olds, but does not differ between the samples for other age groups (Figure 5).

Voluntary Activity

Our dependent variable is a binary indicator of whether or not the respondent reported having done any voluntary work in the past year. Simple survey measures of voluntary work are believed to have good reliability properties and to have similar measurement properties across subgroups, even though they may not have high validity (Lynn 1994). We used the same predictor variables and the same model-fitting technique as for vote turnout. Our final model included as predictors political interest, age, education, marital status and activity status. We found significant interactions of sample with both political interest and age. For those who report being “not at all interested in politics” the predicted probability of taking part in voluntary work is significantly lower in the mixed mode sample than in the face-to-face sample, whereas the same is not true of respondents who report being “very”, “quite” or “hardly” interested in politics (Figure 6). As regards age, predicted probabilities for 15-24

year-olds are lower in the mixed mode sample , while the pattern across the older age groups is similar with both mode protocols (Figure 7).

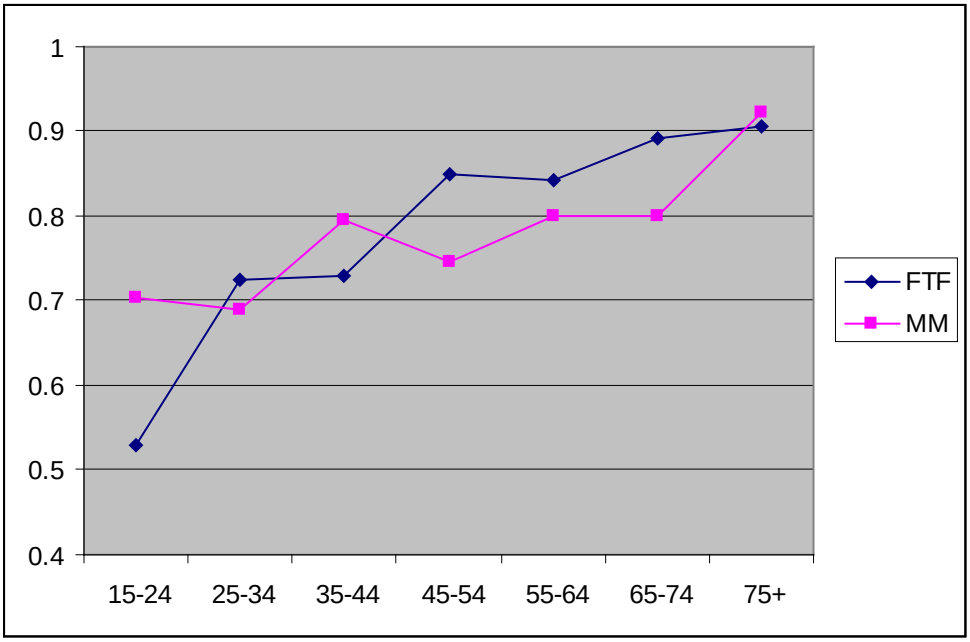
Broad Political Participation

Respondents were asked whether or not they had taken each of seven forms of action in the last 12 months, namely contacting a politician or government official, working in a political party or action group, working in another organisation or association, wearing or displaying a campaign badge or sticker, signing a petition, taking part in a lawful public demonstration, and boycotting certain products. We model a binary indicator of whether or not the respondent reported having taken at least one of these actions (overall, 46% had done so). The final model includes as predictors political interest, gender, age, education and activity status. We find a significant interaction of sample with gender: predictions differ between samples only for females (Figure 8).

Active Political Participation

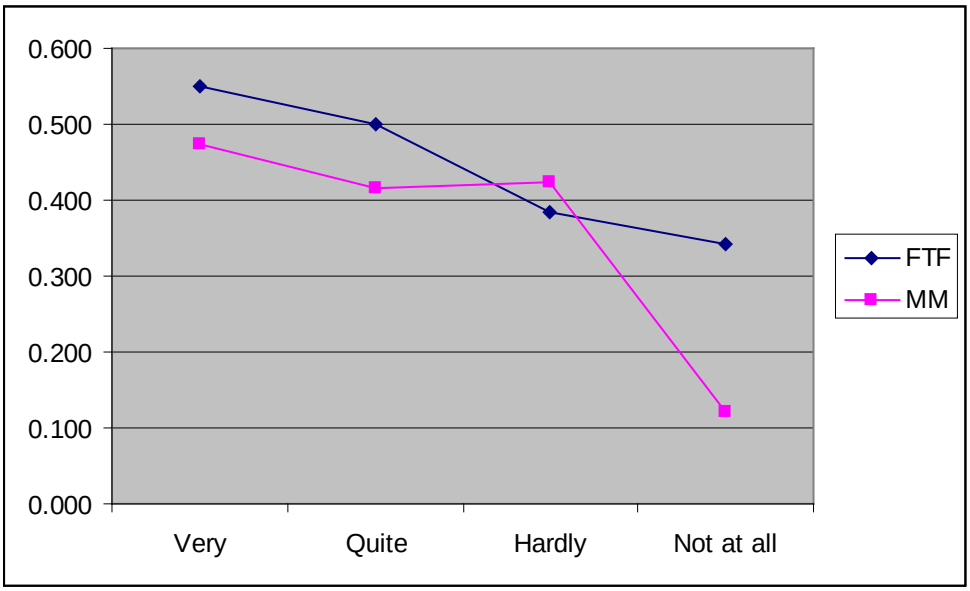
Of the seven activities included in our measure of broad political participation, two - displaying a campaign badge or sticker and signing a petition - can be considered somewhat passive activities, which take little effort and may not indicate the same level of involvement as the other activities. We therefore develop an alternative indicator which does not include these two activities and therefore reflects a more active participation. Overall, 38% of the sample had taken part in at least one of the other five activities. The final model includes as predictors political interest, gender, education, marital status and activity status. We find a significant interaction of sample with gender: women were significantly less likely to report political participation in the face-to-face sample, while for men there was no difference between the samples (Figure 9).

Figure 5: Predicted Probability of Turnout: Interaction between Age and Mode



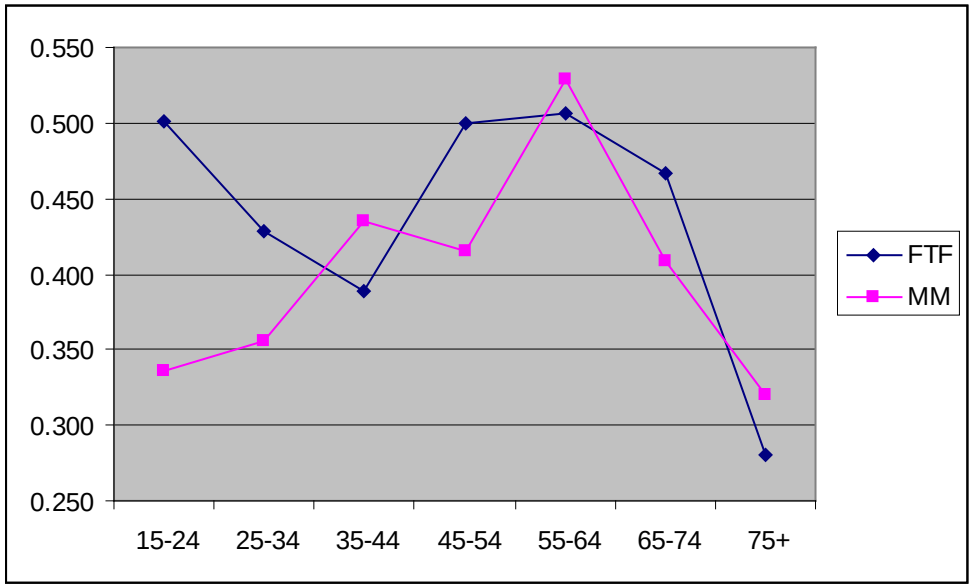
Note: Y-axis shows predicted probabilities of self-reported vote for a married male in paid work, with medium education and who reports being "quite interested" in politics.

Figure 6: Predicted Probability of Voluntary Work: Interaction between Political Interest and Mode



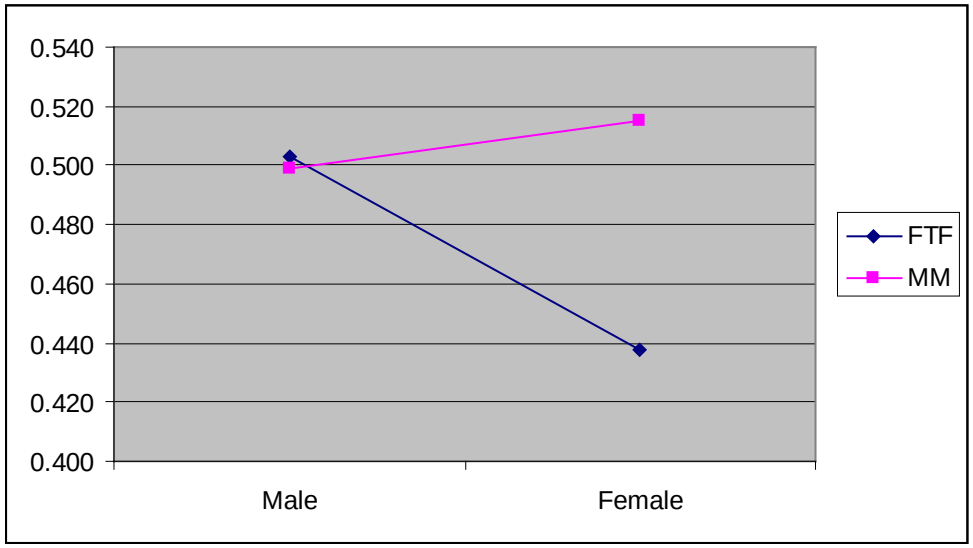
Note: Y-axis shows predicted probabilities of having done voluntary work in the past year for a married male aged 45-54 in paid work, with medium education.

Figure 7: Predicted Probability of Voluntary Work: Interaction between Age and Mode



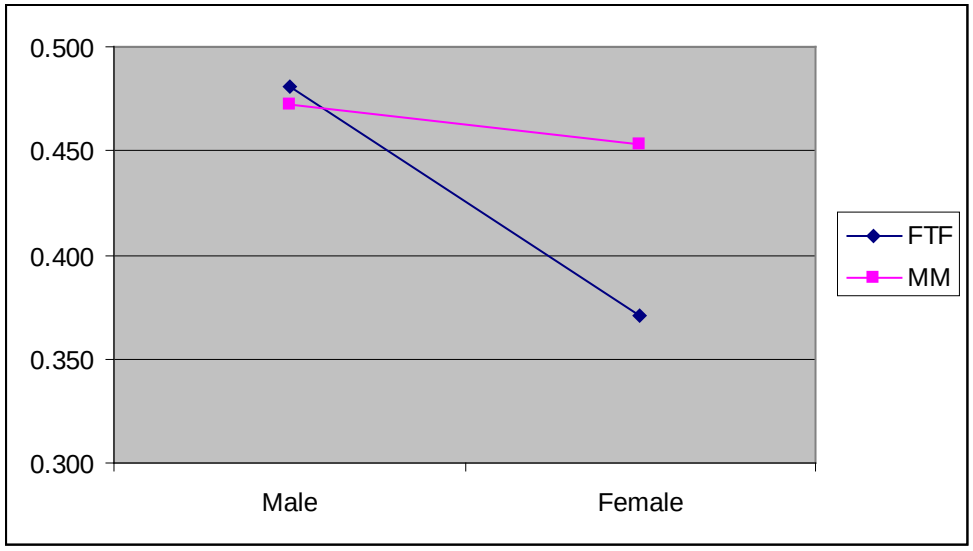
Note: Y-axis shows predicted probabilities of having done voluntary work in the past year for a married male in paid work, with medium education and who reports being "quite interested" in politics.

Figure 8: Predicted Probability of Broad Political Participation: Interaction between Gender and Mode



Note: Y-axis shows predicted probabilities of having done voluntary work in the past year for a 45-54 year-old in paid work, with medium education and who reports being "quite interested" in politics.

Figure 9: Predicted Probability of Active Political Participation: Interaction between Gender and Mode



Note: Y-axis shows predicted probabilities of having done voluntary work in the past year for a married person in paid work, with medium education and who reports being "quite interested" in politics.

6 Conclusions

Our univariate analysis suggests that data collection protocol can affect the distribution of survey variables, though for the majority of items no significant difference was observed between protocols. The items most susceptible to protocol effects appear to be 4-, 5- and 6-point attitude scales and ordinal behavioural items. We find suggestive evidence that some of the differences may be due to reduced social desirability bias with web interviews and greater stated pessimism with web interviews. Our investigation of scale equivalence, on the other hand, reveals little evidence against the hypothesis of measurement invariance between protocols. This suggests that the correlational structure of related items may not be disturbed by the data collection protocol, even if the marginal distributions of the items are affected.

However, estimation of a number of statistical models involving an indicator of political interest reveals some modest differences between protocols in the fitted models. This indicates that estimated multivariate relationships between variables might differ depending on whether the survey data are collected by face-to-face or mixed mode interviewing. The evidence from the modelling of political interest is not overwhelming. We found that in the mixed mode sample, respondents who value material equality are more likely to be politically

interested than those who do not, whereas in the face-to-face sample, the opposite appears to be true. However, since only two out of sixteen examined variables showed a significant interaction effect with data collection protocol, it is possible that simple chance has produced the observed differences between the samples. (The probability to obtain two statistically significant results by chance would be 4.3 % if we examined 16 predictors that are independent of one another; since the two variables D1 and D4 are not independent of one another, the actual probability to obtain two significant results may be even higher.)

The evidence is perhaps a little stronger from the models in which political interest is used as a predictor variable. Out of six predictor variables, we find significant interactions with one of them (age) in models of vote turnout, with two of them (age and political interest) in models of voluntary work, and with one of them (gender) in models of political involvement. Even so, the majority of the associations do not differ significantly between the two data collection protocols.

We would also note that in each case where differences in relationships are observed between samples, we do not know which of the two is a more accurate reflection of the true population association, though we do believe that responses to the political interest question will be subject to less social desirability bias with the mixed mode protocol.

We would also note that the precise nature of a protocol effect when comparing a mixed mode design with a face-to-face design will depend on the size and structure of the subsample who respond in each mode. This places limits on the generalisability of our results to other countries, other mixed mode designs, and other points in time. In particular, we suspect, as discussed above, that some of the observed differences in our study may be due to distinct measurement properties when responding b y web. We carried out our study in the Netherlands, which is the ESS country with the second highest level of internet penetration, after Sweden (Eva et al. 2010, p. 6). The numbers and characteristics of people who would respond by web may be different in other countries, potentially leading to different protocol effects.

Despite the caveats about generalisability, this study has indicated that there is good reason to be cautious about comparing findings from surveys carried out with different data collection

protocols – specifically, comparing data from a face-to-face interview survey to data from a mixed mode survey. We would emphasise that protocol effects can impact upon any kind of analysis, are difficult to predict in advance, and are unlikely to be detected – or detectable – in the absence of a detailed methodological study. Even possession of some knowledge about likely effects on univariate distributions does not help us to predict effects on model coefficients or other multivariate estimates.

This has implications for surveys such as the ESS which have relied to date on face-to-face interviewing. Changing to mixed mode data collection may introduce inconsistencies in the time series and make it difficult to identify the nature of change or stability over time. However, more research is needed to understand the reasons for the differences that we have observed. We are urging caution in moving away from face-to-face interviewing not because we necessarily think that data from face-to-face interviews are better, but simply because we think they may be different. The decision facing a survey that has been established in one mode is different from the decision that would face a new survey with no existing time series and no specific need to compare results to previous surveys.

References

Allum N.; Read S.; Sturgis P. (2011) Evaluating social and political trust in Europe. In Davidov E.; Schmidt P.; Billiet J. (eds) *Cross-cultural analysis. Methods and applications*. New York and London: Routledge (pp. 35-53).

Bowling A. (2005) Mode of questionnaire administration can have serious effects on data quality, *Journal of Public Health* 27, 281-291.

Butler D. and Stokes D.E. (1971) *Political Change in Britain*. College Edition. New York: St. Martin's Press.

Chen F., Curran P. J., Bollen K. A., Kirby J., and Paxton P. (2008) An Emprical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods and Research* 36 (4): 462-494.

Christensen A. I., Ekholm O., Juel K. (2011) Comparison of Personal Interviewing and Self-administered Questionnaires: Effect on representativeness and Prevalence of Selected health Indicators. Presentation given at the *Fourth Conference of the European Survey Research Association (ESRA)*. Lausanne, Switzerland, 18-22 July 2011.

Clarke H. D., Sanders D., Stewart M. C. and Whiteley P. F. (2003) Britain (not) at the polls, 2001, *Political Science and Politics* 36, 59-64.

De Leeuw E. D. (2005) To mix or not to mix data collection modes in surveys, *Journal of Official Statistics* 21 (2), 233-255.

Dillman, D. A., Smyth, J.D., and Christian, L. M. (2009) *Internet, mail, and mixed mode surveys: the tailored design method*. Hoboken, NJ: Wiley.

Eva G., Loosveldt G., Lynn P., Martin P., Revilla M., Saris W., and Vannieuwenhuyze J. (2010) ESSPrep WP6 – Mixed Mode Experiment. Deliverable 21: Final Mode Report. London: European Social Survey.

Gabriel O. W. and van Deth J. W. (1998) Political interest, Chapter 14 in The Impact of *Values*, van Deth J W and Scarbrough E (ed.s), Oxford University Press

Halperin E. and Sulitzeanu-Kenan R. (2010) Making a Difference: Political Efficacy and Policy Preferences Polarization. Available at SSRN: http://ssrn.com/abstract=1719564

Kreuter F., Presser S. and Tourangeau R. (2008) Social Desirability Bias in CATI, IVR, and Web Surveys: the Effects of Mode and Question Sensitivity, *Public Opinion Quarterly* 72, 847-865.

Lynn P. (1994) Measuring voluntary activity, *Non-Profit Studies*, 1 (2), 1-11.

Lynn P., Uhrig S.C.N. and Burton J. (2010) Lessons from a randomised experiment with mixed mode designs for a household panel survey, *Understanding Society Working Paper* 2010-03, Colchester: University of Essex.

Martin P. (2010) "I hope I'm not a racist". The investigation of everyday racism using surveys. *Unpublished PhD Dissertation. Department of Sociology, City University, London*. July 2010.

Martin P. (2011) What makes a good mix? Chances and challenges of mixed mode data collection in the ESS. Centre for Comparative Social Surveys Working Paper No. 02. London: Centre for Comparative Social Surveys, City University. Available online:

http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=855&itemid=80 .

Messer B.L. and Dillman D.A. (2009) Improving the effectiveness of mail contact procedures to obtain survey response over the internet for general public household surveys, paper presented at the 64th *Conference of the American Association for Public Opinion Research*, Hollywood, Florida.

Meuleman B. & Davidov E. (2008) *European attitudes towards immigration, 2002 - 2007. A cross-country and cross-time comparison*. Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC) . Berlin, Germany June 25-28, 2008. Available: http://www.csdiworkshop.org/pdf/3mc2008_proceedings/session_82/Meuleman_(Bart)_Davidov.pdf

Meuleman, B. & Billiet, J. (2011) Religious involvement: its relation to values and social attitudes. In: Davidov E.; Schmidt P.; Billet J. (eds) *Cross-cultural analysis. Methods and applications*. New York: Routledge (pp. 173-206).

Revilla M. (2010) "Quality in Unimode and Mixed-Mode designs: A Multi-trait Multimethod approach". *Survey Research Methods* 4(3): 151-164.

Roberts C., Jäckle A., & 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 of the American Statistical Association, 4221-4228*

Schwartz S. H. (2007a) Value orientations: measurement, antecedents and consequences across nations. In: Jowell, R; Roberts, C; Fitzgerald, R & Eva, G (eds) *Measuring attitudes cross-nationally. Lessons from the European Social Survey*. Los Angeles: Sage (pp. 169-203).

Schwartz S.H. (2007b) Universalism: Values and the Inclusiveness of Our Moral Universe, *Journal of Cross-Cultural Psychology* 38 (6), 711-728

Swaddle K. and Heath A. (1989) Official and reported turnout in the British general election of 1987, *British Journal of Political Science* 19, 537-551.

Tourangeau R. and Yan, T. (2007) Sensitive Questions in Surveys, *Psychological Bulletin.* 133 (5), 859-883.

Van de Vijver F. J. R. (2011) Capturing Bias in Structural Equation Modeling. In: Davidov E.; Schmidt P.; Billet J. (eds) *Cross-cultural analysis. Methods and applications*. New York: Routledge (pp. 3-34).

Vannieuwenhuyze J., Loosveldt G. and Molenberghs G. (2010) A Method for Evaluating Mode Effects in Mixed-mode Surveys, *Public Opinion Quarterly* 74(5), 1027-1045

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 (3), 367-387.

Annex A: Variables used in Modelling the Determinants of Political Interest

Concept	ESS variable	Survey questions	Scale
Sex	F2_01		binary
Age	age	(derived variable from year of birth)	continuous
Education	edulvl	Derived from F6	1-3 (High, Medium, Low)
Religiosity	C21	Regardless of whether you belong to a particular religion, how religious would you say you are?	0-10 (10: Very religious)
	C22	Apart from special occasions such as weddings and funerals, about how often do you attend religious services nowadays?	1-7 (1: Every day; 7: Never)
	C23	Apart from when you are at religious services, how often, if at all, do you pray?	1-7 (1: Every day; 7: Never)
Left-right materialism	B30	The government should take measures to reduce differences in income levels.	1-5 (1: Agree Strongly)
	D1	Large differences in people's incomes are acceptable to properly reward differences in talents and efforts.	1-5 (1: Agree Strongly)
	D4	For a society to be fair, differences in people's standard of living should be small.	1-5 (1: Agree Strongly)
New egalitarianism	B31	Gay men and lesbians should be free to live their own life as they wish	1-5 (1: Agree Strongly)
	D3	A woman should be prepared to cut down on her paid work for the sake of her family.	1-5 (1: Agree Strongly)
	D6	When jobs are scarce, men should have more right to a job than women.	1-5 (1: Agree Strongly)
Other questions	B32	Political parties that wish to overthrow democracy should be banned.	1-5 (1: Agree Strongly)
	D2	Schools must teach children to obey authority.	1-5 (1: Agree Strongly)
	D5	People who break the law should be given much harsher sentences than they are these days.	1-5 (1: Agree Strongly)
	B23	In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right?	0-10 (0: Left; 10: Right)

Annex B: Variables used in Modelling the Determinants of Voting, Voluntary Activity and Political Involvement

Concept	ESS variable	Survey questions	Scale
Sex	F2_01		binary
Age	age	(derived variable from year of birth)	continuous
Education	edulvl	Derived from F6: What is the highest level of education you have achieved?	1-3 (High, Medium, Low)
Marital status		Derived from F62	3 categories (married or civil partnership, never married or partnered, formerly married or partnered)
Economic activity status		Derived from f8cmainact	5 categories (paid work, education, unemployed or disabled, retired, other)
Turnout	B11	Some people don't vote nowadays for one reason or another. Did you vote in the last Dutch national election in November 2006?	binary
Voluntary activity	E49	In the last month have you done any paid or voluntary work?	binary (voluntary work only or paid work and voluntary work *vs*. paid work only or neither)
Broad political involvement		Derived from B13 to B19	binary (yes to at least one of B13-B19 *vs*. none)
Active political involvement		Derived from B13, B14, B15, B18, B19	binary (yes to at least one of B13, B14, B15, B18, B19 *vs*. none)
	B13	Have you contacted a politician or government official in the last 12 months?	binary
	B14	Have you worked in a political party or action group in the last 12 months?	binary
	B15	Have you worked in another organisation or association in the last 12 months?	binary
	B16	Have you worn or displayed a campaign badge/sticker in the last 12 months?	binary
	B17	Have you signed a petition in the last 12 months?	binary
	B18	Have you taken part in a lawful public demonstration in the last 12 months?	binary
	B19	Have you boycotted certain products in the last 12 months?	binary