ELABORATED EVALUATION OF MODE EFFECTS IN MIXED MODE DATA

!Submitted to journal, not for quotation

Authors:
Vannieuwenhuyze Jorre T. A. & Loosveldt Geert
Katholieke Universiteit Leuven, Belgium

Correspondence:
e-mail: jorre.vannieuwenhuyze@soc.kuleuven.be
Adress: Jorre Vannieuwenhuyze, OE Centrum voor Sociologisch Onderzoek, Parkstraat 45 - bus 3601, 3000 Leuven, België
tel : +32 16 323144
fax : +32 16 323365

words: 7007

Abstract: Survey designs in which data from different groups of respondents are collected by different survey modes, become increasingly popular. However, such a mixed-mode (MM) designs lead to a confusion of selection effects (response bias) and measurement effects (measurement bias) caused by mode differences. Consequently, MM data has poor quality. Nevertheless, comparing MM data with data from a comparable single-mode survey allows measuring selection effects and measurement effects separately. The authors illustrate how to evaluate mode effects using data from a Dutch MM experiment within the European Social Survey program. In this experiment, respondents could choose between three modes: a web survey, a telephone interview, or a face-to-face interview. Mode effects on three political variables are evaluated: interest in politics, perceived complexity of politics and voter turnout in the last national election.

Keywords: Measurement Error, Mode effects, Mixed-Mode, Nonresponse error, Selection effects

1. Introduction

Increasingly, data are gathered by mixing different survey modes in one design (Weisberg, 2005). Dillman (2009) distinguishes 4 types of mixed-mode (MM) designs of which one refers to the collection of the same data from different sample members by different modes. Such a MM data collection can help reduce coverage error, lower nonresponse and nonresponse bias, and reduce costs (de Leeuw, 2005; Don A. Dillman et al., 2009). Sample coverage may be ameliorated because several modes are available to contact different groups of hard-to-reach respondents. Response may be augmented since every respondent can choose his mode of preference between several modes. Costs may be reduced because a substantial part of the sample will be surveyed by a cheap mode.

However, the question is whether MM designs, notwithstanding their advantages, automatically lead to higher data quality (Voogt & Saris, 2005). MM designs may lower nonresponse bias and avoid coverage error, but they may introduce other forms of bias as well. Mode effects can make MM data highly unusable by simultaneously generating nonresponse error (selection effects) and measurement error (measurement effects). This article illustrates a method to disentangle these two types of error when comparable single-mode data is available.

2. Selection effects and measurement effects

Selection effects occur when different types of respondents choose different modes to complete the survey. As such they are forms of nonresponse error, i.e. different types of respondents do not respond in certain modes. The occurrence of a selection effect is in itself not a problem. On the contrary, its occurrence makes using a MM design valuable. Indeed, because of selection effects, some respondents may accept participation while they would not in a single-mode survey. It has been recognized that some respondents prefer being surveyed by one mode while others prefer a different mode (Biemer, 2001; Day et al., 1995; de Leeuw & Van Der Zowen, 1988; D. A. Dillman et al., 2009; Voogt & Saris, 2005). Similarly, others will accept participation by a cheap mode lowering total survey costs.

Measurement effects refer to the influence of a survey mode on the answers respondents give, such that one person would give different answers in different modes (Bowling, 2005; Voogt & Saris, 2005; Weisberg, 2005, p. 278). Put differently, measurement effects are

caused by differences in measurement errors (Groves, 1989). These errors may originate from differences in, among others, whether items are presented sequentially or simultaneously to the respondent, interviewer effects and social desirability, primacy and recency effects, recall bias, acquiescence, etc. (Bowling, 2005; Brick & Lepowski, 2008; de Leeuw, 1992; de Leeuw, 2005; D. A. Dillman, 1991; Don A. Dillman et al., 2009; Schwarz et al., 1991).

The major problem of MM designs is that selection effects and measurement effects are inextricably bounded. Differences (or similarities) between the outcomes of modes can be caused by differences between the respondents or by differences in measurement error (de Leeuw, 1992; Weisberg, 2005). Exclusive focus on MM survey data precludes evaluation of selection effects and measurement effects separately. However, this problem is avoided by an extension of MM data with data of a comparable single-mode survey. The next section will present the dataset used to illustrate the evaluation of mode effects. Section 4 addresses how mode effects can be calculated while section 5 will focus on drawing inferences on mode effects.

3. Data

3.1. ESS and mixed mode experiment

The European Social Survey (ESS) started in 2002 as a biennial survey conducted in 30 European countries. Its goal is to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. It contains topics like trust, politics, social values, social exclusion, discrimination, religion, national identity, life course, … So far, four waves of data gathering have been performed with the last wave fielded in 2008/2009.

In order to encourage equivalence across countries, all ESS surveys have completely been carried out by face-to-face (FTF) personal interviews so far. Because of the expensiveness of FTF interviews, declining funds, declining response rates, changing coverage issues, and the resistance from certain countries without a tradition of conducting FTF interviews, a Mixed-Mode (MM) experiment was set up parallel to the 4th round in the Netherlands.

The purpose of this MM experiment was to compare a mixed mode survey design with the main Dutch ESS survey by using exactly the same questionnaire. In this study, we will use the 2674 FTF sample members of the main Dutch ESS data with matched telephone numbers.

These respondents were reached by at most 10 interviewer contact attempts at home. In the MM experiment, a sample of 878 persons with a matched phone number was drawn from the Dutch population and assigned to a concurrent MM design[1]. In contrast to the main ESS survey, sample members of the MM experiment could choose between three survey modes, a web questionnaire (CAWI), a telephonic interview (CATI) or a face-to-face personal interview at home (CAPI). The survey started with a first telephonic contact (1st telephonic screening) including 14 call attempts. If a person was willing to participate the survey, the different survey modes were offered simultaneously so that the respondent could immediately choose his or her preferred mode. All sample members who could not be contacted or refused to participate in the 1st screening were subject to a 2nd telephonic screening. This 2nd screening was performed analogously to the 1st screening.

The follow up of nonresponse depended on the mode someone chose in the telephonic screenings. First, the respondents who chose to complete the web questionnaire were recontacted at most 14 times telephonically to remind them to complete the questionnaire. If a respondent refused to complete the web questionnaire, still a telephonic or FTF interview was offered.

Second, the sample members who chose a telephonic interview were either interviewed immediately during the telephonic screening or an appointment was made for a call back. Although these sample members were allowed changing their mind and asking for a web survey or a FTF interview, only one switched to a Web survey. Nonresponse could occur if there was no contact at an appointment. These nonrespondents were approached FTF for a personal interview in a follow up phase after the telephonic screening phase.

[1] The MM expereriment contains in fact 2500 sample members. 744 of these could not be matched to a telephone number. These sample members were approached for a FTF interview at home, if they refused other survey modes were offered. However the majority of these respondents (>85%) completed the survey by CAPI, which makes this group unattractive to evaluate mode effects. The remaining 1756 sample members were divided between a sequential and concurrent MM design (878 each). In the sequential design the modes were offered sequentially (first web, then telephone, then FTF) in stead of simultaneously. Unfortunalty, the quality of the implemented sequential design and the data can be doubted. As such we only restrict our research to the concurrent MM data.

Lastly, the respondents who chose a FTF interview were visited by an interviewer at home. Non-contacts or nonresponse were not followed up in an other survey mode.

Sample members who could not be contacted or who refused to participate during the telephonic screening were subject to a FTF follow up as well. These respondents were offered to complete a personal interview. If they refused, still the web survey and the telephone survey were offered, in that order.

Both datasets were separately weighted on a set of socio-demographical variables (age, sex, urbanization, household size, education) increasing the population representativeness.

3.2. Comparability of the ESS round 4 and the MM experiment data

As already been noted, the particular design of the MM experiment involves selection effects (differences in preference for a particular mode) and measurement error caused by measurement effects being inextricably confounded and thus impossible to be evaluated. Differences or similarities between respondents answering in different modes, are caused by an imperceptible interplay between selection and measurement effects. However, comparing the MM data with the ESS round 4 data allows distinguishing the CAPI respondents from the CATI and CAWI respondents and disentangling measurement effects and selection effects, as will be illustrated below.

Nevertheless, when comparing the ESS round 4 data with the mixed mode data, we implicitly assume that the realized samples contain a comparable set of respondents. Comparable samples are samples of respondents of which differences in the distribution of the unbiased version of the variable(s) of interest are completely caused by sampling error. As a consequence, the effect of nonresponse error must be equal in both samples. Unfortunately, this assumption is immediately contradictory to one of the starting points of MM surveys. Indeed, MM survey are considered attractive because they may raise response rates and sample coverage compared to single-mode surveys. On the other side, it is well-known and generally observed that CAPI has an almost perfect coverage and often results in high response rates (relative to the other modes) (de Leeuw, 1992). Consequently, a switch from a single-mode CAPI survey to a mixed mode survey is probably mainly driven by the idea of lowering costs rather than increasing response and coverage.

The comparability assumption can be substantiated by comparing the response rates and the social-demographical composition of the different samples. The response frequencies and response rates can be found in Table 1. While mixed-mode designs are usually used to

increase response rates, the response rate of the ESS MM experiment is, remarkably, significantly smaller than the response rate of the main ESS survey. This inequality is probably caused by differences between the two surveys in efforts made to reach all sample members. However, this is a preliminary conclusion because, by using different modes, the MM design possibly attracted more hard-to-reach CAPI respondents while putting less effort to reach other easy-to-reach respondents. Consequently, the MM sample is possibly comparable with the main ESS round 4 data, even though its response rates are lower.

Further, the comparability can be checked by comparing the composition of both (unweighted) datasets on a set of 'mode-insensitive' socio-demographical variables. The unweighted realized samples of the MM experiment and ESS round 4 were compared on several socio-demographical variables (age, sex, urbanization, household size, education) (tables not included). No significant differences could be found between the ESS round 4 and the MM data. This is an argument enforcing the comparability assumption. However, this argument is not decisive either because it is only valid if these socio-demographical variables are closely related with the unbiased version of the variable(s) of interest. We will return to the issue of comparability in the discussion.

A partial solution to correct for large differences in socio-demographical composition and response rates, is the calculation of propensity scores based on the socio-demographical variables. These propensity scores should be calculated for all respondents of one sample in comparison with the other sample. Afterwards the inverse propensity scores could be used as weights in order to make the two samples comparable. The two samples we use, however, are already weighted for analysis on the very same set of variables to make them better representing the population. As a consequence, both datasets are already comparable on these socio-demographical characteristics.

Table 1: Response frequencies and response rates

	ESS MM exp.	ESS round 4
CAWI	163	
CATI	90	
CAPI	114	1366
total response	367	1366
nonresonse	432	1192
not eligible	79	116
total sample	878	2674
response rate*	45,93%	53,40%

based on sample members with matched phone number only

** = total response/(total sample - not eligible)*

3.3. Variables

It has been argued that measurement effects are strongest on questions about sensitive topics (Brick & Lepowski, 2008; Schwarz et al., 1991; Voogt & Saris, 2005; Weisberg, 2005). These questions are particularly susceptible for one type of measurement error, namely measurement error caused by social desirability in presence of an interviewer (Aquilino, 1994; Bowling, 2005; Don A. Dillman, 2005; D. A. Dillman et al., 2009; Voogt & Saris, 2005). Because of the present interaction between interviewer and respondent, respondents act by social norms and give cultural acceptable answers. In face-to-face surveys, people tend to over report behaviour that is considered socially desirable, while this tendency occurs less frequently in self-reported questionnaires (de Leeuw, 1992; Voogt & Saris, 2005; Weisberg, 2005).

In this article we will start with the analysis of the variable political interest. Respondents were asked how interested they are in politics and could choose one out of four answer categories: (1) not at all interested, (2) hardly interested, (3) quite interested, and (4) very interested. The higher the score, the more interest. Political interest is considered particularly susceptible to social desirability. It is assumed that people tend to over report their interest in presence of an interviewer because politics (or interest in politics) is seen as a civic duty (Voogt & Van Kempen, 2002).

In the next two sections, political interest will be used to illustrate the methods to evaluate mode effects. Afterwards, in the 6^{the} section, we will look whether the observed mode effects can also be found on other politics related variables: perceived political complexity and voter turnout. Respondents were asked how often politics seem so complicated that they can't really understand what is going on. Five possible answers were offered: (1) never, (2) seldom, (3) occasionally, (4) regularly, and (5) frequently. Higher scores refer to more complexity. Perceived complexity of politics is highly correlated with political interest (e.g. in the ESS round 4 the correlation = -0.433, $p<0.001$). Highly interested person generally evaluate politics less complex. As a consequence, we expect mode effects on this variable as well. Further, the respondents were asked whether they voted in the last Dutch national election in November 2006, yes (1) or no (2). Voters are usually more interested in politics (difference in interest = 0.673 between voters and nonvoters in the ESS round 4, $p<0.001$) and voting is a social desirable behaviour.

4. Method

Before expounding the general idea behind the method to evaluate measurement effects and selection effects, we need to establish some symbolic notations which will be used throughout this paper. First, we denote with I the variable 'interest in politics' (1= not at all interested, 2=hardly interested, 3 =quite interested, 4 =very interested). However, we will not use I for analysis, but we will rather consider two (sub)variables I_P and I_{WT}. With I_P, we refer to political interest measured by CAPI, I_{WT}, on the other hand, refers to political interest measured by CATI or CAWI. Considering the outcome of different survey modes as different variables allows us to evaluate measurement effects just by comparing I_P and I_{WT} for the same respondents. Off course, given the survey design either I_P or I_{WT} is observed for each respondent. So, this problem should be circumvented. The variable I_P has a multinomial distribution with parameter vector $\boldsymbol{\pi}_{\mathbf{P}} = (\pi_{P1}, \pi_{P2}, \pi_{P3}, \pi_{P4})$ where π_{Pi} is the chance that $I_P = i$, $0 \le \pi_{Pi} \le 1$ and $\sum_i \pi_{Pi} = 1$ for $i = 1,2,3,4$. The distribution of I_{WT} is defined analogously.

M is the mode the respondent 'chooses' when he or she is or would be a respondent of the Mixed-Mode experiment. In principle, this variable has three categories, but, since we can only compare CAPI with a combination of CATI and CAWI, we reduce the form of M to a binomial variable with values 0 for the choice of CATI or CAWI and 1 for CAPI. M has one distributional parameter τ, $0 \le \tau \le 1$, which represents the chance that a respondent chooses CAPI in the Mixed Mode experiment.

Given the available data we can estimate several distributions:

- $P(I_P)$ from the ESS round 4 data, which is a sample completely surveyed by CAPI.
- $P(I_P | M = 1)$ from the MM data, more specifically from the respondents who responded by CAPI in the MM experiment.
- $P(I_{WT} | M = 0)$ from the MM data as well, but now from the respondents who responded by CATI or CAWI in the MM experiment.
- $P(M = 1)$ and $P(M = 0)$ from the whole MM data set

Knowing this information we can apply the following formula. By the law of total probability

$$P(I_P) = P(I_P|M=0)P(M=0) + P(I_P|M=1)P(M=1).$$

After some rearrangement of the terms we get

$$P(I_P|M=0) = P(I_P)\frac{1}{P(M=0)} - P(I_P|M=1)\frac{P(M=1)}{P(M=0)} \qquad (2.1)$$

The left part of equation (2.1) is the distribution of political interest when this question would have been asked by CAPI, for respondents who actually chose CATI or CAWI in the MM-experiment. Since these respondents answered by CATI or CAPI and not by CAPI we can not derive this distribution from the data directly. However, all distributions on the right side of the equation can be derived.

Since $P(I_P = i) = \pi_{Pi}$, we can use equation (2.1) to reconstruct the probabilities $\pi_{Pi}|M=0$, the probabilities of the four categories of I_P for the respondents choosing CATI or CAWI. Afterwards we can compare these probabilities with

- the observed probabilities of the respondents who chose CAPI in the MM experiment. Differences between these result from selection effects because the probabilities represent the distribution of the same variable (political interest measurement by CAPI) for different groups of respondents (different mode preference). We define the *selection effect* as $(\pi_{Pi}|M=1) - (\pi_{Pi}|M=0)$, the probability for the CAPI choosing respondents minus the probability for the CATI or CAWI choosing respondents.
- the observed probabilities of the respondents who chose CATI or CAWI. Differences between these are caused by measurement effects since we compare outcomes of the same group of respondents (with a preference for CAWI or CATI) from different modes. We define the *measurement effect* as $(\pi_{WTi}|M=0) - (\pi_{Pi}|M=0)$, the probability of the political interest scores measured by CATI or CAWI minus the probability measured by CAPI.

A summary of this strategy to evaluate measurement effects and selection effects can be found Table 2. Figure 1 gives a hypothetical graphical representation of this strategy. The bar plots are a representation of a distribution of political interest scores measured by CAPI (I_P,

left barplot) or by CAWI or CATI (I_{WT}, right barplot). The total height of these bars represent the relative frequency of the 4 possible answer categories. The black contoured bars in the left plot represent the distribution in the main ESS round 4 survey which is fully conducted by CAPI, i.e. $P(I_P)$. The grey bars represent the distribution of I_P for the respondents in the Mixed Mode experiment who answered by CAPI, i.e. $P(I_P|M=1)$. Finally, the black bars in the right plot give a picture of the distribution for the respondents of the MM experiment who chose CAWI or CATI, i.e. $P(I_{WT}|M=0)$. Note that the Y-axis refers to a 'relative' rather than the 'absolute' frequency. Since we want to compare the ESS round 4 with the Mixed Mode experiment, the surface of the black contoured bars should be equal to the sum of the surface of the grey and black bars. This means that the sample sizes should be equal, which can easily be obtained by data weighting.

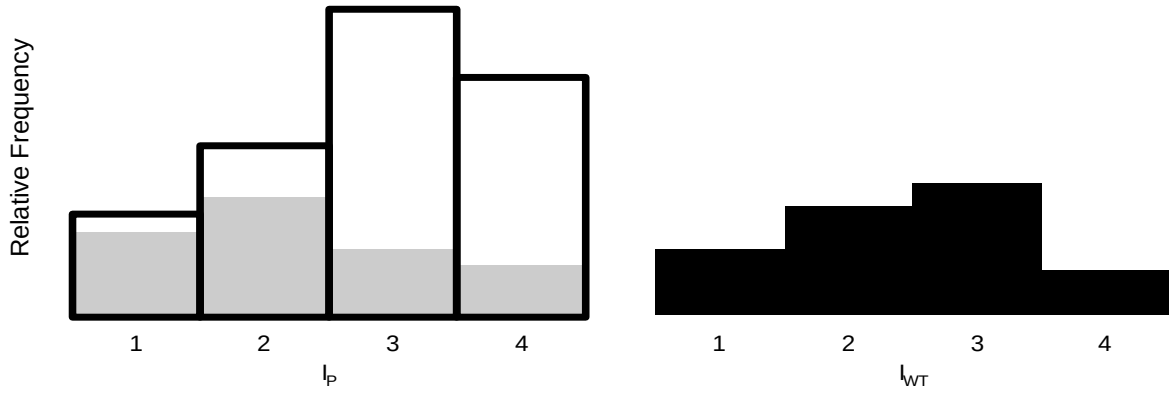
As a consequence, the remaining white surface of the 4 bars in the left barplot must correspond with the distribution of $I_P|M=0$. This is the political interest registered by CAPI of persons who would choose a CAWI or CATI when they are selected for the MM experiment. We can now compare these white areas with the black areas in the right bar plot. The differences between these areas point at measurement effects or differences among measurement error between modes. Since the white surfaces in the I_P bar plot represents $P(I_P|M=0)$ and the black surfaces in the I_{WT} bar plot represent $P(I_{WT}|M=0)$, they represent the distribution of the same variable (political interest) for the same group of persons (those who would chose CAWI or CATI in the MM experiment) but actually measured by different modes.

Analogously, we can compare the white surfaces with the grey surfaces in the I_P bar plot. These are both distributions of the same variable measured in the same mode, but for different groups of respondents (those who choose CAPI versus those who choose CATI or CAWI in a MM design). As a consequence, differences between these grey and white surfaces must represent selection effects.

Table 2: Overview of statistics to evaluate selection effects and measurement effects

statistic	derivation	area in Figure 1
π_P	from ESS round 4 data	black contours
$\pi_P \mid M=1$	from MM exp, CAPI group	grey
$\pi_{WT} \mid M=0$	from MM exp, CATI and CAWI group	black
τ	from total MM exp	
$\pi_P \mid M=0$	$=\pi_P/(1-\tau)-(\pi_P \mid M=1)\tau/(1-\tau)$	white
selection effect	$=(\pi_P \mid M=1)-(\pi_P \mid M=0)$	grey-white
measurement effect	$=(\pi_{WT} \mid M=0)-(\pi_P \mid M=0)$	black-white

Figure 1: Graphical representation of hypothetical distributions of the data to illustrate the strategy to disentangle measurement effects from selection effects



The total height of the black bordered bars in the barplot of I_P represent the relative frequencies of the political interest scores in the ESS round 4 (i.e. $P(I_P)$). The grey areas represent the frequencies of the interest scores of the respondents in de MM experiment who answered by CAPI ($P(I_P|M=1)$). The barplot of I_{WT}, represents the distribution of the respondents of the MM experiment who answered by CAWI or CATI ($P(I_{WT}|M=0)$).

In Table 3 the reader can find the sample probabilities observed in the data, as well as the calculated probabilities $\pi_{Pi} \mid M=0$, the measurement effects and selection effects based on the formulas in Table 2. In terms of observed percentages, the CAPI respondents in the MM experiment (column 3) chose the answer categories 'quite interested' and certainly 'not at all interested' more frequently than their CATI or CAWI choosing colleagues (column 2). The categories 'hardly interested' and 'very interested', on the other hand, are relatively more chosen by the CAWI/CATI respondents. Remind that differences between these probabilities can be due to selection effects as well as measurement effects.

Using the ESS round 4 data, we partially can evaluate selection effects and measurement effects, which are shown in the last two columns of Table 3. As these columns make clear, some selection effects and measurement effects are noticeable. With respect to the category 'hardly interested' the measurement effect and selection effect are positive. With respect to the other categories, they are negative, except for the selection effect of 'not at all interested'. First, this means that more respondents will indicate to be 'hardly interested' when this question is asked by CAWI or CATI compared to the situation when this question is asked by CAPI. This trend is at the expense of the number of respondents indicating the other three categories in CAWI or CATI compared to CAPI. Second, because the selection effect is positive, the respondents who choose CAPI in the MM experiment are more likely to choose the category 'not at all interested' and 'hardly interested' compared to the respondents choosing CAWI or CATI. These latter are more likely to indicate the other two categories.

The interpretation of the probabilities of each answer category and their accompanying mode effects, is rather long-winded. To make things easier to interpret, we combined the parameter vector $\boldsymbol{\pi}_P$ into one parameter, the mean μ_P which reflects the mean interest in politics. The mean can simply be calculated by applying the formula,

$$\mu_P = 1*\pi_{P1} + 2*\pi_{P2} + 3*\pi_{P3} + 4*\pi_{P4}$$

to the observed and calculated parameters. In the same way, the variance can be calculated as well by the formula

$$\sigma_P^2 = (1-\mu_P)^2 * \pi_{P1} + (2-\mu_P)^2 * \pi_{P2} + (3-\mu_P)^2 * \pi_{P3} + (4-\mu_P)^2 * \pi_{P4}.$$

Analogously we can calculate μ_{WT} and σ_{WT}^2. Once these means and variances are calculated, we can evaluate measurement effects and selection effects on these parameters once again. Measurement effects on the mean political interest indicates that one survey mode over- or underestimates the overall political interest in comparison with the other mode. Measurement effects on the variance indicate that the answers are more heterogeneous for one mode in comparison with the other mode. This possibly has an influence on inferences about several statistics.

The means and variances with their measurement effects and selection effects can be found in Table 3 as well. The measurement effect as well as the selection effect on the mean political interest are negative. Apparently, a CAPI mode measures a higher mean political interest compared to a combination of CAWI and CATI, which translates in a measurement

effect. This supports the hypothesis that respondents over report their interest in politics in front of an interviewer because it is social desirable behaviour. On the other side, the negative selection effect on the mean interest tells us that the respondents choosing CAPI in the MM experiment are on average less interested in politics than their CATI or CAWI choosing colleagues. This might be explained by the observation that respondents who choose CAPI in a MM survey, are often nonrespondents in a single-mode CATI or CAWI survey. These nonrespondents are usually less interested in politics (Voogt & Van Kempen, 2002).

The measurement effect on the variance is negative as well. This indicates that a CAPI survey results in more heterogeneous answers compared to a combination of CATI and CAWI. Respondents are more likely to give comparable answers in the latter survey modes. This might be explained by a primacy or recency effect in a telephone interview which involves that respondents tend to answer the first or the last offered answer category. As a result, their answers are more homogeneous. The selection effect on the variance is very small (but positive). The respondents choosing CATI or CAWI answer more or less as homogeneously as the CAPI choosing respondents.

Table 3: Estimation of measurement effect and selection effect for political interest

category	$I_{WT}\|M=0$	$I_P\|M=1$	I_P	$I_P\|M=0$	measur. effect	selection effect
P(not at all interested)	0,047	0,109	0,088	0,080	-0,033	0,029
P(hardly interested)	0,389	0,296	0,232	0,207	0,182	0,088
P(quite interested)	0,480	0,535	0,577	0,593	-0,113	-0,058
P(very interested)	0,084	0,060	0,103	0,120	-0,036	-0,060
mean	2,601	2,547	2,696	2,753	-0,152	-0,206
variance	0,501	0,585	0,594	0,585	-0,084	0,000

5. Inferences for selection effects and measurement effects

The calculated mode effects in the previous section mark a trend, but they do not ascertain whether these effects are significant. In this section, we broaden the method of the previous section taking into account the uncertainty about the measured mode effects, such that inferences can be made.

The sampling distribution of the mode effects can be calculated using frequentist or likelihood approaches. When the sampling variance of these statistics is known, hypotheses about the measurement effect and selection effect can be tested and confidence intervals can be set up. However, in this paper we will use a Bayesian framework to evaluate the mode

effects (for more information about Bayesian data analysis see for example Gelman et al., 2004; Gill, 2002). A Bayesian framework leads to more straightforward calculations and interpretations because a Bayesian point of view implicitly takes into account that the parameters of the distribution of I_P, $I_P|M=1$ and M have a distribution on their own. These parameter distributions reflect the uncertainty about the population parameters of the variables at hand. Keeping in mind this uncertainty, we can easily calculate the parameters of the distribution of $I_P|M=0$, or, more specific, their parametrical distribution. These parametrical distributions can then be used to evaluate the measurement effects and selection effects within a context of uncertainty.

In Bayesian analysis, the uncertainty of the distributional parameters, represented by θ, is included by setting up a probability distribution for these distributional parameters conditional on the observed data y, i.e. $p(\theta|y)$. By Bayes rule, it follows that

$$p(\theta|y)=\frac{p(y|\theta)\,p(\theta)}{p(y)},$$

which is called the posterior distribution of θ given the data. $p(y|\theta)$ is the likelihood function. $p(\theta)$ is a prior distribution, to be specified in advance and usually taken noninformative. $p(y)=\sum_\theta p(y|\theta)\,p(\theta)$ or the total probability of y for all possible values of θ, which is constant with respect to θ. In this case, the posterior distributions will be derived from the data of two samples: the ESS round 4 and the MM experiment data.

For I_P, we like to get the posterior distribution $P(\boldsymbol{\pi}_{\mathbf{P}}|y_{r4})$, where y_{r4} represents the data of the ESS round 4. As a prior distribution for $\boldsymbol{\pi}_{\mathbf{P}}$ we use a Dirichlet distribution with parameters $\alpha_i=1$, $i=1,2,3,4$, which is a commonly used noninformative prior distribution for the parameters of a multinomial distribution. The posterior distribution then is Dirichlet as well:

$$\boldsymbol{\pi}_{\mathbf{P}}|y_{r4}\sim Dirichlet(1+y_1,1+y_2,1+y_3,1+y_4)$$

where y_i is the number of respondents who answered category i in the ESS round 4. The posterior distribution of $\boldsymbol{\pi}_{\mathbf{P}}|M=1,y_{MM}$ and $\boldsymbol{\pi}_{\mathbf{WT}}|M=0,y_{MM}$, where y_{MM} represents the data of the MM experiment, can be obtained analogously.

With respect to M, we model $P(\tau | y_{MM})$. A commonly used noninformative prior distribution is $\tau \sim U[0,1]$. This prior results in a posterior distribution $\tau | y_{MM} \sim Beta(y+1, n-y+1)$ with y the number of CAPI respondents in the dataset, and n the total number of respondents (in y_{MM}).

An approach often used in Bayesian analysis, is to simulate/take random draws from the components at the right hand side of the equation (Gelman et al., 2004; for simulation techniques we refer to Ripley, 1987; and Robert & Casella, 1999). $\boldsymbol{\pi}_{\mathbf{P}} | M=0, y_{r4}, y_{MM}$ can then be calculated for each simulation and the combination of these simulations represents $P(\boldsymbol{\pi}_{\mathbf{P}} | M=0, y_{r4}, y_{MM})$. Afterwards, the measurement effects and selection effects can be calculated by subtracting simulated values of $\boldsymbol{\pi}_{\mathbf{P}} | M=0, y_{r4}, y_{MM}$ from simulated values of $\boldsymbol{\pi}_{\mathbf{WT}} | M=0, y_{MM}$ and $\boldsymbol{\pi}_{\mathbf{P}} | M=1, y_{MM}$ respectively. As a result we get a set of simulated values of the measurement effects and selection effects, reflecting their posterior distribution conditional on the data. Using these sets of simulations we can easily set up confidence intervals and make inferences about the mode effects.

We made 10.000 simulations of $\boldsymbol{\pi}_{\mathbf{P}} | M=0, y_{r4}, y_{MM}$, the measurement effects and the selection effects. We also recalculated the mean political interest score and the variance for each simulation, and their accompanying measurement effect and selection effect. The calculation of 10.000 simulations allows us to estimate p-values up to four decimal places (the definition of a p-value will be given below). The results can be found in Table 4.

The second column of Table 4 contains the median of the 10.000 simulations of the measurement effects and the selection effects. These medians can be used to make overall statements about these effects. However, it is a wise idea to first make a graphical check to get an idea about the shape of the distributions. All distributions are shown in Figure 2. Apparently, these distributions are more or less symmetric with the majority of the simulations lying in the centre and a minority in the tails. As a result, the median can be used as a good summarizing statistic of this distributions. Off course, the medians are in line with the conclusions made in the previous section. These conclusion will not be repeated.

Of higher importance, are the last three columns of Table 4. These columns tell us something about the uncertainty of measuring the measurement effects and selection effects. The third column shows the $2{,}5^{th}$ percentile of the simulated measurement effects/selection

effects. This means that 2,5% of all simulated values of the measurement effects/selection effects are lower than this percentile. Analogously, the fourth column contains the 97,5th percentile. The combination of column three and four is a 95% confidence interval. 95% of the simulated values fall between the 2,5th percentile, the lower bound, and the 97.5th percentile, the upper bound. This confidence interval can be used to make inferences about the measurement effects and the selection effects. If zero falls outside this confidence interval, we say that a measurement effect or selection effect is significant at the 5% level.

With respect to the measurement effects, there is a clear effect on the probability of indicating 'hardly interested' and 'quite interested'. If a respondent answers in CATI or CAWI, it is more likely that he will choose the answer category 'hardly interested' than when he answers in CAPI. The opposite is true for the category 'quite interested'. No significant measurement effects can be noticed for the two other answer categories. The negative measurement effect on the mean political interest is significant as well. Respondents will on average report a higher political interest when the survey is conducted by CAPI than in a survey conducted by CATI or CAWI. By contrast, no significant measurement effect is found on the variance. Our data suggest that the results of a CAPI survey and a survey conducted by CAWI and CATI, are equally heterogeneous.

Considering the selection effects, no significant effects can be found. All 95% confidence intervals contain zero, the value which indicates at an absence of a selection effect. As a consequence, there is no statistical evidence that respondents who choose CAPI in a MM design are different from respondents who choose CATI or CAWI with respect to their political interest.

The last column of Table 4 contains a p-value. This p value corresponds to the proportion of simulated measurement effects/selection effects with a sign opposite to the median simulated value. If the median is positive, this p is the proportion of negative simulated measurement effects/selection effects. If the median is negative, it is the proportion of positive simulations. In contrast to the 95% confidence interval, which implies a 2-sided test, this p-value can act as a one sided test for the measurement effects or selection effects. The interpretation of these p-values is the probability that a measurement effect or a selection effect is positive or negative.

These p-values are somewhat less conservative than the 95%-confidence intervals. By definition, all mode effects which are significant by the 95%-confidence interval, are certainly

significant based on the p-values. However, the opposite reasoning is not true. So, it's worth to look at effects which become significant by the one-sided test. The selection effect on the mean political interest is significantly negative in both MM designs. So, although zero falls within the central 95% confidence interval, there is strong evidence for a negative selection effect on the mean political interest. This means that respondents who choose CAPI in a MM design, are less interested in politics compared with their CATI or CAWI choosing colleagues.

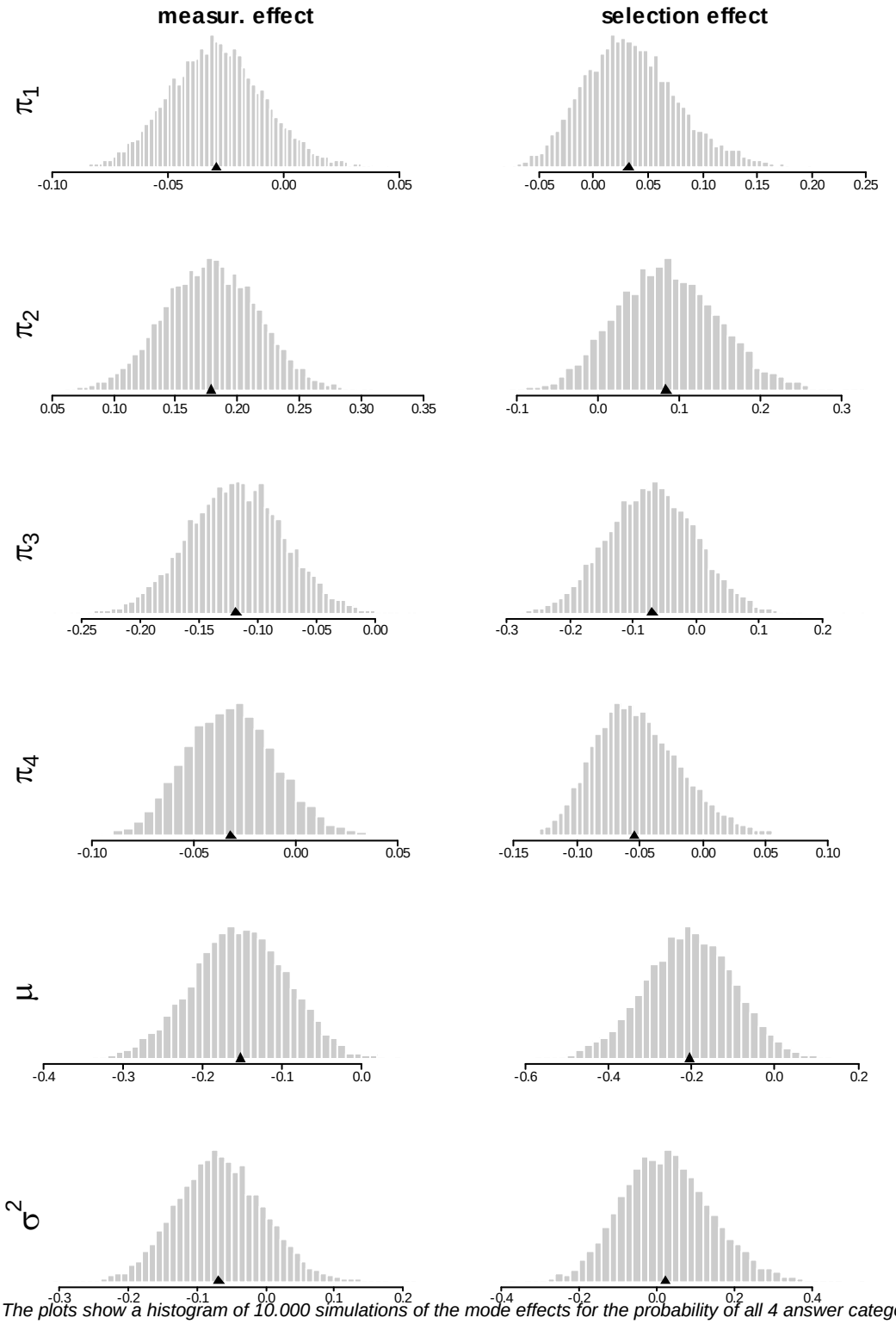
Table 4: Inferences for measurement effects and selection effects for political interest in the mixed mode designs

		95% confidence interval		
	median	2,5 percentile	97,5 percentile	p*
MEASUREMENT EFFECT				
P(not at all interested)	-0,029	-0,068	0,014	0,087
P(hardly interested)	0,179	0,106	0,253	0,000
P(quite interested)	-0,119	-0,198	-0,040	0,002
P(very interested)	-0,032	-0,074	0,015	0,081
mean	-0,153	-0,273	-0,038	0,005
variance	-0,068	-0,187	0,063	0,150
SELECTION EFFECT				
P(not at all interested)	0,032	-0,041	0,131	0,219
P(hardly interested)	0,083	-0,033	0,218	0,086
P(quite interested)	-0,070	-0,205	0,068	0,158
P(very interested)	-0,054	-0,111	0,027	0,086
mean	-0,206	-0,425	0,001	0,026
variance	0,023	-0,185	0,271	0,425

based on 10.000 simulations

** The probability p refers to the number of simulations with a measurement effect/selection effect smaller than zero (P(sim<0)) for positive median values, or larger than zero (P(sim>0)) for negative median values.*

Figure 2: Histograms of simulated measurement effects and selection effects



The plots show a histogram of 10.000 simulations of the mode effects for the probability of all 4 answer categories and the mean. The black triangles at the bottom of the plots mark the medians.

6. Political complexity and voter turnout

In the previous sections, it became clear that some measurement effects and selection effects are noticeable on the distribution of political interest. In this section we will repeat the methods used to evaluate measurement effects and selection effects on two other politics related variables, perceived complexity and voter turnout in the last national election. As before 10.000 simulated scores of the measurement effects and the selection effects were drawn.

In Table 5, the results of perceived political complexity can be found. Considering a two-sided test at the 95% confidence level, there seems only to be a significant measurement effect on the probability of the answer category 'seldom', which is negative. Respondents are thus less likely to consider politics seldom complex when they answer this question by CATI or CAWI compared to the situation when they answer by CAPI. Only two significant selection effects are found, namely on the probability of selecting the answer 'never' and on the variance (the $97{,}5^{th}$ percentile is slightly negative). Respondents choosing the CAPI mode are less likely to never find politics too complex and their answers are less heterogeneous, than respondents choosing CATI or CAWI. Considering the one-sided test for the sign of selection or measurement effects, the former conclusions can be broadened with a significantly positive measurement effect on the mean. So, some significant measurement effects and selection effects can be found with respect to the variable political complexity.

Table 6 summarizes the results of the variable voter turnout. Since this variable has only two answer categories, measurement effects and selection effects are complementary for both probabilities (did vote or did not vote). Table 6 only shows measurement effects and selection effects on the probability that someone did vote in the last national election and the variance of this variable. No measurement effects significantly different from zero can be noticed. Apparently, a combination of CATI and CAWI as survey modes seems to result in the same probability of voting compared to a survey totally conducted by CAPI. The p-values of the variance are equal to the p-values of the probability. This is logical, since the outcome of each respondent is a result of a Bernoulli trial with parameter τ and variance equal to $\tau(1-\tau)$, a function of the parameter. Consequently, we test twice the same measurement effect or selection effect resulting in the same p-value. A significant selection effect, on the contrary, is found and this effect is negative. The respondents who chose CAPI are less likely to have

voted in the last election, compared to the other respondents. This could be expected since these type of respondents are less interested in politics, as was shown before.

Table 5: Estimated measurement effects and selection effects for perceived political complexity

		95% confidence interval		
	median	2,5 percentile	97,5 percentile	p*
MEASUREMENT EFFECT				
P(never)	-0,001	-0,041	0,043	0,477
P(seldom)	-0,083	-0,142	-0,020	0,006
P(occasionally)	0,037	-0,039	0,115	0,167
P(regularly)	0,025	-0,040	0,094	0,233
P(frequently)	0,020	-0,028	0,071	0,218
mean	0,149	-0,025	0,319	0,048
variance	0,004	-0,220	0,228	0,486
SELECTION EFFECT				
P(never)	-0,072	-0,110	-0,010	0,013
P(seldom)	-0,031	-0,133	0,091	0,302
P(occasionally)	0,078	-0,052	0,218	0,125
P(regularly)	0,025	-0,085	0,150	0,333
P(frequently)	-0,013	-0,078	0,079	0,376
mean	0,173	-0,087	0,441	0,100
variance	-0,349	-0,647	0,000	0,025

based on 10.000 simulations

** The probability p refers to the number of simulations with a mode effect/selection effect smaller than zero (P(sim<0)) for positive median values, or larger than zero (P(sim>0)) for negative median values.*

Table 6: Estimated measurement effects and selection effects for voter turnout

		95% confidence interval		
	median	2,5 percentile	97,5 percentile	p*
MEASUREMENT EFFECT				
P(voted)	-0,020	-0,076	0,030	0,226
variance	0,016	-0,024	0,061	0,226
SELECTION EFFECT				
P(voted)	-0,116	-0,238	-0,015	0,011
variance	0,083	0,011	0,157	0,011

based on 10.000 simulations

** The probability p refers to the number of simulations with a measurement effect/selection effect smaller than zero (P(sim<0)) for positive median values, or larger than zero (P(sim>0)) for negative median values.*

7. Discussion

The purpose of this article is to illustrate how two different types of mode effects, i.e. selection effects and measurement effects, can be disentangled within a MM survey context. Given a MM survey dataset, it is impossible to evaluate both types of error separately.

However, we showed that the presence of data from a single-mode comparative survey allows measuring selection effects and measurement effects and making inferences about them.

As an illustration, we tried to detect selection effects and measurement effects in a mixed mode survey experiment within the ESS. In this experiment a sample was offered the choice to complete the questionnaire by three different modes, a web questionnaire, a telephone interview or a face-to-face interview. We compared data from this mixed mode experiment with the data from the main ESS which is fully conducted by face-to-face interviews. Consequently, we could disentangle measurement effects from selection effects between the CAPI mode on the one hand and the CATI and CAWI mode on the other hand. We evaluated mode effects on the variables political interest, perceived complexity of politics and voter turnout in the last national election.

In this article, we did find some measurement effects and selection effects within the ESS data. Measurement effects point at differences between CAPI surveys and CATI or CAWI surveys with respect to the level of measurement bias. Selection effects point at differences between the respondents who choose or respond by CAPI in a mixed mode design and respondents who respond by CATI or CAWI. However, we like to add three important caveats.

First, the reader should bear in mind that the measurement effects are not calculated on the whole sample, but on a part of the sample. We calculated the measurement effect by the difference between the statistics obtained in a CAPI survey and a CATI/CAWI survey respectively, only for the respondents who choose CATI or CAWI in a MM design. The question is whether these measurement effects can be generalized to the respondents who choose the CAPI mode in the MM survey.

Second, in order to measure the different mode effects we explicitly assumed the MM data (MM experiment) and the single-mode comparative data (main ESS) to be comparable. Two samples are comparable if differences between the distributions of the unbiased version of the variables of interest are only caused by sampling error. This means that the effect of nonresponse error or bias must be equal in both samples. This is possibly a strong assumption and certainly the weakest part of the method we used. The comparability assumption is immediately contradictory with some starting points of MM designs, namely to increase coverage and nonresponse. However, in the situation that a single-mode FTF survey is compared with a MM survey, like the ESS, it can be questioned whether this is an issue. It is

generally observed that FTF surveys have high coverage and relatively high response rates. Respondents are more likely to participate in CAPI than CATI or CAWI surveys. It is natural to assume that respondents who accept participation in a web or telephone survey would accept participation in a FTF survey as well. The main question is which proportion of these CAWI and CATI respondents would really participate in a single-mode CAPI survey like the ESS round 4. If this proportion is high (towards 100%) and the same efforts are made in both samples with respect to the CAPI interviews (i.e. follow up, number of contacts, …), one can easily assume that both samples are comparable. We suggest further research towards mode preferences and most importantly 'mode acceptance' of respondents. With mode acceptance we refer to the willingness of respondents to participate in a particular mode, regardless whether this is his most preferred mode.

It is quasi impossible to evaluate the comparability assumption given the specific design of the ESS data. All we can do is making a good guess. Two arguments can be postulated to substantiate the assumption. On the one hand, the two samples can be compared with respect to the response rates. If the response rates are comparable and rather high, one could conclude that the samples are comparable as well. However, this argument is weak. Especially in a MM context, lower response rates do not necessarily imply a lower representativeness and lower efforts to reach the respondents. Indeed, by using different modes it might be easier to include respondents who are hard-to-reach in a single-mode survey. As such, the representativeness of a MM design might be equal to or better than a single-mode design, although the response rate is lower.

On the other hand, the distribution of some socio-demographical variables can be compared among the two samples. Data is often compared with national census data in order to evaluate the representativeness of the sample. Large differences between these socio-demographical distributions can be used as an argument against comparability of the samples. However, this socio-demographical is not soul-saving either. Indeed, it assumes that these socio-demographical variables are completely mode-insensitive and highly correlated with the unbiased (true) version of the variable(s) of interest.

Third, attention must be paid to the conclusions to be drawn with respect to the ESS data. The method we offered works fine when the MM data is gathered by only two modes. However, the specific design of the ESS survey and the MM experiment with 3 modes prevents us evaluating mode effects in full detail.

On the one hand, if significant measurement effects and selection effects are found, these measurement effects and selection effects are present. However, these effects only embody a difference between CAPI and a combination of CATI and CAWI where each respondent can choose between the latter two survey modes. Given the design, these latter two survey modes are inextricably bounded in the conclusions, and it is completely impossible to measure differences in mode effects between CAWI and CATI given the ESS data. Consequently, the obtained measurement effects do not inform about the difference between CAPI and CATI, nor between CAPI and CAWI respectively. A measurement effect can be caused by CATI or CAWI separately, or by an interplay of these two modes. In the same way, significant selection effects do not necessarily mean that CATI respondents differ from CAPI respondents, nor that CAWI respondents differ from CAPI respondents, and if they differ in what direction this difference is.

On the other hand, if no significant measurement effects or selection effects are found, interpretations become even less straightforward. Indeed, the absence of significant effects does not allow us to conclude that there are no measurement effects or selection effects at all. The mode effects and selection effects specific to CATI and CAWI respectively, may counteract. As a result the overall effect of a combination of both modes may become small.

The ESS data and the MM experiment made clear how measurement effects can be disentangled from selection effects if a comparative single-mode group is available. However, the specific character of the data, i.e. the presence of three survey modes in the MM experiment and only one mode in the comparative dataset, prevents us from measuring the mode effects between all modes exactly. This problem would not have happened if the MM experiment counted only two modes (CAPI and an other mode). We therefore suggest including at most two modes in future MM designs, if only one comparative single-mode sample is available. The first mode should be the mode of the comparative group, the second mode can be any other survey mode. As such, exact differences in measurement error or bias can be evaluated between these two modes. If one wants to add more modes to a MM survey, more comparative groups should be available.

References:

Aquilino, W. S. (1994). Interview Mode Effects in Surveys of Drug and Alcohol-Use - a Field Experiment. *Public Opinion Quarterly, 58*(2), 210-240.

Biemer, P. P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics, 17*(2), 295-320.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291.

Brick, J. M., & Lepowski, J. M. (2008). Multiple Mode and Frame Telephone Surveys. In J. M. Lepowski, L. Japec, C. Tucker, P. J. Lavrakas, J. M. Brick, M. W. Link, E. D. De Leeuw & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 149-169). Hoboken: Wiley & Sons.

Day, N. A., Dunt, D. R., & Day, S. (1995). Maximizing Response to Surveys in Health Program Evaluation At Minimum Cost Using Multiple Methods. *Evaluation Review, 19*(4), 436-450.

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

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

de Leeuw, E. D., & Van Der Zowen, H. (1988). Data quality in telephone and face-to-face surveys: a comparative analysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 283-299). New York: Wiley-Interscience.

Dillman, D. A. (1991). The Design and Administration of Mail Surveys. *Annual Review of Sociology, 17*, 225-249.

Dillman, D. A. (2005). Survey mode as a source of instability in responses across surveys. *Field methods, 17*(1), 30-52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research, 38*(1), 3-20.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail and Mixed-Mode Surveys : the Tailored design method.* Hoboken: Wiley.

ESS Round 4: European Social Survey Round 4 Data. (2008). Data file edition 1.0.: Norwegian Social Science Data Services, Norway - Data Archive and distributor of ESS data.

Gelman, A., Carlin, J. S., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton (Fla.): Chapman and Hall/CRC.

Gill, J. (2002). *Bayesian methods : a social and behavioral sciences approach*. Boca Raton (Fla.): Chapman and Hall/CRC.

Groves, R. M. (1989). *Survey errors and survey costs*. New York (N.Y.): Wiley.

Ripley, B. D. (1987). *Stochastic simulation*. New York (N.Y.): Wiley.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York (N.Y.): Springer.

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The Impact of Administration Mode on Response Effects in Survey Measurement. *Applied Cognitive Psychology, 5*(3), 193-212.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics, 21*(3), 367-387.

Voogt, R. J. J., & Van Kempen, H. (2002). Nonresponse bias and stimulus effects in the Dutch National Election Study. *Quality & quantity, 36*(4), 325-345.

Weisberg, H. F. (2005). *The total survey error approach : a guide to the new science of survey research.* . Chicago (Ill.): University of Chicago.