A Method for Evaluating Mode Effects in Mixed-mode Surveys

Vannieuwenhuyze, Jorre; Loosveldt, Geert & MOLENBERGHS, Geert (2010) A Method for Evaluating Mode Effects in Mixed-mode Surveys. In: PUBLIC OPINION QUARTERLY, 74 (5). p. 1027-1045.

DOI: 10.1093/poq/nfq059
Handle: http://hdl.handle.net/1942/11843

ELABORATED EVALUATION OF MODE EFFECTS IN MIXED MODE DATA

Authors:

Vannieuwenhuyze Jorre T. A.[1], Loosveldt Geert[1], and Molenberghs Geert[2]

[1]: Centre of Survey Methodology, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

[2]: l-BioStat, Katholieke Universiteit Leuven, B-3000 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: 5658

Abstract:

Survey designs in which data from different groups of respondents are collected by different survey modes, become increasingly popular. However, such mixed-mode (MM) designs lead to a confounding of selection effects (nonresponse error) and measurement effects (measurement error) 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 Survey, Selection effects, European Social Survey

1. Introduction

Increasingly, data are gathered by mixing different survey modes in one design (Don A. Dillman et al., 2009; Weisberg, 2005). One type of such mixed-mode (MM) designs includes the collection of the same data from different sample members by different modes. Such a MM data collection can help reduce coverage error, and lower nonresponse and nonresponse bias in order to reduce the Total Survey Error (TSE), or it can help reduce costs (de Leeuw, 2005; Don A. Dillman et al., 2009). Sample coverage may be improved 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, notwithstanding their advantages, MM designs do not automatically lead to higher data quality or smaller TSE (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).

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. various types of respondents do not respond in certain modes by self-selecting themselves for another mode. 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 (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, on the other hand, 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).

In order to evaluate the TSE introduced by a MM data collection, selection effects and measurement effects should be investigated separately. The major problem of MM designs, however, is that selection effects and measurement effects are completely confounded. 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). The literature suggests using response matching on a set of mode-insensitive variables (e.g. gender, age, education level, etc.) to disentangle both mode effects (e.g. de Leeuw, 2005). Nevertheless, this method assumes that the matching variables are closely related with the variables of interest, but this assumption can mostly be doubted. So, exclusive focus on MM survey data almost precludes evaluation of selection effects and measurement effects separately.

However, comparing MM data with data of a comparable single-mode survey allows disentangling mode effects to a certain extent. This article aims to illustrates a method to disentangle measurement effects from selection effects by comparing a MM dataset with a comparable single-mode dataset. The next section will present the dataset used to illustrate this evaluation of mode effects. Section 3 addresses how mode effects can be evaluated on a multinomial variable, while section 4 reports the estimated mode effects in the data at hand.

2. Data

2.1. 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, among others, trust, politics, social values, social exclusion, discrimination, religion, national identity, and 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 MM experiment was set up in the Netherlands parallel to the 4th round.

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 which could be matched to a telephone number in the sampling list. 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 very same sampling list and assigned to a concurrent MM design. In this concurrent design, sample members could choose between three survey modes, a web questionnaire (CAWI), a telephonic interview (CATI) or a face-to-face personal interview at home (CAPI), from the very first contact[1]. Sample members without matched telephone were also included in both the main ESS and the MM experiment, but almost all of the experiment respondents responded by a FTF interview as well. Consequently, this group is hardly useful to evaluate mode effects.

Both samples contain a random draw of households in which one household member older than 15 years was selected randomly. To correct for differences in household sizes, normalized design weights proportional to the household size were used in all analyses.

[1] The MM expereriment also contains a sequential design in which modes were offered sequentially (first web, then telephone, then FTF) in stead of simultaneously. However, we only restrict our analyses to the concurrent MM data.

The MM experiment started with a 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 second telephonic screening. This second screening was performed analogously to the first 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. Nonetheless, these nonrespondents were not automatically recontacted by an interviewer at their house.

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 to change their mind and to ask 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.

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

Response frequencies of both datasets can be found in Table 1. For convenience, respondents with partial incomplete answers on the variables described in the next section were left out for the further analyses. Both the main ESS data and the MM experiment data were further separately weighted on a set of socio-demographical variables (age x sex, urbanization, and household size) increasing the population representativeness. The marginal population distributions of these variables were obtained from the 'Centraal bureau voor de statistiek (CBS)[2]'. The adjusting post-stratification weights were calculated using iterative proportional fitting or raking procedures (Deming & Stephan, 1940; Izrael et al., 2000).

To end we should make one additional remark. The MM sample is gathered by three survey modes and selection effects and measurement effects can be expected between all of these modes. However, as the method in section 3 will illustrate, comparison between the single-mode FTF main ESS and the MM experiment only allows evaluating differences between CAPI on the one hand and a combination of CATI and CAWI on the other hand. The latter two modes cannot be compared with each other without additional assumptions. This problem would not have happened if the MM experiment contained only 2 modes (CAPI and any other mode).

[Table 1 about here.]

[2] *www.cbs.nl*

2.2. Variables

In this article we will separately analyse three politics-related variables: political interest, perceived political complexity, and voter turnout. 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. Subsequently, respondents were asked how often politics seems so complicated that they cannot really understand what is going on. Five possible answers were offered: (1) never, (2) seldom, (3) occasionally, (4) regularly, and (5) frequently. Further, the respondents were asked whether they voted in the last Dutch national election in November 2006, yes (1) or no (2).

In the CAPI mode all answer categories were read out to the respondent by the interviewer in the right order (reversed order as mentioned above for political interest), excluding “don’t know”-categories. For the political complexity question, the reading was accompanied by a showcard with all five substantial answer categories. In the CATI mode, the question and the answers were read out to the respondent analogous to CAPI but no showcards were used. In the CAWI mode the questions were shown using the very same wording and order of answer categories. If the respondent tried to skip a question, however, a ‘don’t know’ answer appeared at the bottom of the answer list. The respondent was obliged to select one answer.

All of the three variables are expected to be susceptible to mode effects. First, political interest may be affected by a measurement effect because it is seen as a civic duty (Voogt & Van Kempen, 2002). It has been argued that measurement effects are strongest on questions about such socially desirable behaviour (Brick & Lepowski, 2008; Schwarz et al., 1991; Voogt & Saris, 2005; Weisberg, 2005). Because of the present interaction between interviewer and respondent, respondents act by social norms and give cultural acceptable answers in an interview survey. As a consequence, we expect that people tend to over report their interest in face-to-face surveys, while this tendency will occur less frequently in self-reported questionnaires (Aquilino, 1994; Bowling, 2005; de Leeuw, 1992; Don A. Dillman, 2005; D. A. Dillman et al., 2009; Voogt & Saris, 2005; Weisberg, 2005). Perceived complexity of politics and voter turnout can be assumed to be highly correlated with political interest (e.g. in the ESS round 4 the correlation between interest en perceived complexity is -0.433, p<0.001; the difference in interest is 0.673 between voters and nonvoters in the ESS round 4, p<0.001). Highly interested people generally evaluate politics less complex and voters are usually more interested in politics. As a consequence, we expect mode effects on these variables as well.

Second, Voogt & Van Kempen (2002) also argue that nonrespondents are usually less interested in politics. As a consequence, because the CAPI group of MM experiment contains a considerable group of nonrespondents of the first phase of the survey, we can expect selection effects on all three variables. We expect that the CAPI choosers of the MM experiment are less interested in politics, perceive politics as more complicated, and are less likely to have voted in the last election.

Below we will denote with I the variable of interest (political interest, political complexity, or voter turnout). However, we will not use I for analysis, but we will rather consider two (sub)variables I_P and I_{WT}. By I_P, we refer to the variable measured by CAPI, I_{WT}, on the other hand, refers to the variable measured by CATI or CAWI. Considering the outcome of different survey modes as different variables allows us to evaluate measurement effects merely by comparing I_P and I_{WT} for the same respondents. Of course, given the survey design either I_P or I_{WT} is observed for each respondent. So, this problem should be circumvented.

Additionally, we define variable M as 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 binary variable with values 1 for the choice of CAPI and 2 for CATI or CAWI.

3. Method

3.1. Comparability assumption

As already noted in the introduction, we will compare the main ESS data with the MM experiment data to evaluate mode effects. However, in doing so 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 (or purely random nonresponse error). As a consequence, the effect of *systematic* nonresponse error on the variables of interest must be equal in both samples. Unfortunately, this assumption is immediately contradictory to one of the starting points of MM surveys because these surveys are considered attractive to raise response rates and sample coverage compared to single-mode surveys. Nevertheless, two arguments for this '*comparability assumption*' can be investigated.

First, 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. Put differently, we can assume that the CAWI and CATI choosers of the MM experiment, would also accept to participate by a FTF survey when they were sampled for the main ESS round 4 data collection. As a consequence, we expect the response rates of both samples to be comparable. However, the response rate of the ESS MM experiment is, remarkably, significantly smaller than the response rate of the main ESS survey (±7%, see Table 1). This inequality is probably caused by differences between the two surveys in efforts made to reach all sample members. Sample members of the MM experiment who choose to participate by CAWI but did not respond were not followed up by a CAPI indeed. This inaccuracy in sample design might explain the difference in response rates. On the other hand, however, in using different modes the MM design possibly attracted more hard-to-reach CAPI respondents while putting less effort into reaching other easy-to-reach respondents. Consequently, the MM sample may possibly be comparable with the main ESS round 4 data, even though its response rates are lower. So, a comparison of response rates as an argument for the comparability assumption is not decisive.

A second argument for the comparability assumption involves a comparison of the composition of both datasets on a set of 'mode-insensitive' socio-demographical variables. A comparison of the realized samples of the MM experiment and ESS round 4 on several socio-demographical variables (age x sex, urbanization, household size, education) and only corrected by the design weights, however, did not show any significant difference (tables not included). So, this can be used as an argument enforcing the comparability assumption. Still, 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. Nevertheless, we corrected for the small remaining differences using normalized propensity scores weights derived from the complete set of variables mentioned above (Rosenbaum & Rubin, 1983; Sato & Matsuyama, 2003). As a consequence, both datasets are comparable on these socio-demographical characteristics.

3.2. Mode effect calculation

To illustrate the methods applied further, we provide a hypothetical graphical representation of our strategy in Figure 1. The bar plots represents distributions of, for example, political interest scores measured by CAPI (I_P, left barplot) or by CAWI or CATI (I_{WT}, right barplot) and the total height of the bars represent the proportion of respondents choosing one of the 4 possible answer categories in a particular mode of response. Let 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. $P\left(I_P\right)$. The total surface of these bars should equal 1. Let the grey bars represent the distribution of I_P for the respondents in the Mixed Mode experiment who answered by CAPI ($P\left(I_P, M=1\right)$), and the black bars in the right plot represent the distribution for the respondents of the MM experiment who chose CAWI or CATI ($P\left(I_{WT}, M=2\right)$). Once again, the total surface of the grey and black bars should equal 1.

As a consequence, the remaining white surface of the 4 bars in the left barplot must correspond with the distribution of $P\left(I_P, M=2\right)$. This is the political interest registered by CAPI of people who would actually 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 because they represent the distribution of the same variable (political interest) for the same group of people (those who would choose CAWI or CATI in the MM experiment) but 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.

[Figure 1 about here.]

We can now define the selection effect on the proportion parameter of category i as

$$\Sigma_i = P\left(I_P = i \middle| M = 1\right) - P\left(I_P = i \middle| M = 2\right), \tag{1.1}$$

and the measurement effect as

$$\mathrm{M}_i = P\left(I_{WT} = i \middle| M = 2\right) - P\left(I_P = i \middle| M = 2\right). \tag{1.2}$$

$P\left(I_p = i \middle| M = 1\right)$ and $P\left(I_{WT} = i \middle| M = 2\right)$ can simply be estimated with the MM data. $P\left(I_p = i \middle| M = 2\right)$ however is never estimated directly. Nonetheless we can use the law of total probability to prove that

$$P\left(I_P = i \middle| M = 2\right) = P\left(I_P = i\right)\frac{1}{P\left(M = 2\right)} - P\left(I_P = i \middle| M = 1\right)\frac{P\left(M = 1\right)}{P\left(M = 2\right)}. \tag{1.3}$$

If we substitute (1.3) into (1.1) and (1.2), we get:

$$\Sigma_i = \frac{1}{P\left(M = 2\right)}\left[P\left(I_P = i \middle| M = 1\right) - P\left(I_P = i\right)\right], \tag{1.4}$$

and

$$\mathrm{M}_i = P\left(I_{WT} = i \middle| M = 2\right) - P\left(I_P = i\right)\frac{1}{P\left(M = 2\right)} + P\left(I_P = i \middle| M = 1\right)\frac{P\left(M = 1\right)}{P\left(M = 2\right)}. \qquad (1.5)$$

Given the available data we can estimate the distributions of the factors on the right hand side of both (1.4) and (1.5):

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

So, we defined the selection effect and the measurement effect as a function of estimable proportions. Using the fact that the distribution of the sample proportions is asymptotically normal, the delta method can be used to prove that the expected value of the sample mode effects estimates are equal to the population mode effects and that the sample estimates follow an asymptotical normal distribution as well (Agresti, 2002; Casella & Berger, 2002). Further, the delta method also allows deriving approximate estimates of the sampling variance of both effects so that inferences can be built around the estimated mode effects. The exact application of the delta method can be found in the appendix.

An additional question which should be asked when evaluating the mode-effects is whether the sample size n is sufficiently large to detect small to medium effect sizes. The total sample size is the sum of the MM experiment sample size n_M and the main ESS round 4 sample size n_C. It can be shown that n_M and n_C affect the variance of the estimated mode effects by the following relation:

$$\operatorname{var}(effect) = \frac{a_C}{n_C} + \frac{a_M}{n_M}, \tag{1.6}$$

where a_C and a_M are specific functions of the observed sample proportions, but independent from both sample sizes (Refer to the appendix for the calculation of a_C and a_M). Given that

$$\operatorname{var}(effect) = \left(\frac{|effect|}{|Z_{\beta,\alpha}|} \right)^2, \tag{1.7}$$

formula (1.6) can be used to calculate the required sample sizes to achieve a decent power given the critical significance level used. In this equations, $Z_{\beta,\alpha}$ refers to the required Z-value to get a power β given the significance level α, and *effect* is the minimal effect the researcher wants to detect.

The next section gives the results of the estimated selection effects and measurement effects on the proportions of all categories of the three politics-related variables. If these proportions are known, however, mode effects on the mean can be easily calculated as well since the mean is a linear function of the proportions. In contrast to the proportions, the mean allows for straightforward interpretations of the mode effects. The calculation of the mode effects on the mean can be found in the appendix as well. The section concludes with a calculation of required sample sizes to detect moderate mode effects on the means.

4. Results

Table 2 summarizes the observed sample proportions and means which are used to calculate the mode effects estimates. The mean perceived political complexity already shows a remarkable trend. If there were no measurement effects, we could expect that the mean in the main ESS data falls between the means of the two MM groups provided that the comparability assumption holds. The data, however, show a different trend. The mean political complexity in the main ESS is smaller than in both MM groups which might be explained by mode-effects.

[Table 2 about here.]

4.1. Political interest

In Table 3 the reader can find the estimated measurement effects and selection effects for political interest. As this table makes clear, significant measurement effects can be found for the categories ‘hardly interested’ and ‘quite interested’. The measurement effect on the category ‘hardly interested’ is positive which 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. As the measurement on the category ‘quite interested’ is negative, the opposite conclusion is true. Further, the measurement effect on the mean is negative as well, and this is in line with our expectation that the CAPI mode measures a higher mean political interest compared to a combination of CAWI and CATI. As a consequence the one-sided p-value can be used and this turns out to be significant as well. So, respondents may report a higher interest in politics in front of an interviewer because this probably is socially desirable behaviour (Voogt & Van Kempen, 2002).

If the two-sided p-values of the selection effects are considered, none of the selection effects seems to be significant. We expected that the CAPI choosers in MM design were less interested in politics because this group contains more nonrespondents of the first phase of survey (Voogt & Van Kempen, 2002). This means that we expect that the selection effect on the mean is negative, but, as Table 3 shows, this expectation is not met. Consequently, we can not conclude that the respondents choosing CAPI in the MM experiment are on average less interested in politics than their CATI or CAWI choosing colleagues because the former group contains more hard-to-reach respondents.

[Table 3 about here.]

4.2. Perceived political complexity

Table 4 summarizes the estimated mode effects on perceived political complexity. Considering the proportions of all answer categories, there is a significant negative measurement effect on the category ‘seldom’. So, respondents are more likely to consider politics seldom complex when they answer this question by CAPI compared to the situation when they answer by CATI or CAWI. Further the selection effects on ‘never’ and ‘seldom’ are significantly negative, and on ‘occasionally’ significantly positive. So, respondents choosing the CAPI mode, are less likely to never or seldom but more likely to occasionally find politics too complex than respondents choosing CATI or CAWI.

Because we expected the CAPI mode to measure a lower perceived political complexity compared to the CATI/CAWI combination, the measurement effect on the mean should be positive, which is confirmed by the data. Moreover, the one-sided p-value shows that this measurement effect is significant. So, respondents tend to report that they better understand politics when they are surveyed by a personal FTF interview. This observation might be explained by social desirability bias. The sign of the selection effect on the mean comes up to our expectations as well, because a positive selection effect means that the CAPI choosers evaluate politics as more complex. This selection effect is significant as well which confirms our hypothesis.

[Table 4 about here.]

4.3. Voter turnout

Table 5 summarizes the sample proportions and the estimated mode effects 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) and the mean. No measurement effect or selection effect significantly different from zero can be noticed. As a consequence, a combination of CATI and CAWI as survey modes does not seem to result in a different estimation of the probability of voting compared to a survey totally conducted by CAPI. Analogously, a difference in voting behaviour between CAPI choosers and CATI/CAWI choosers is not confirmed either.

[Table 5 about here.]

4.4. Sample size calculation

The sample estimates of a_C and a_M can also be found in Table 3, Table 4, and Table 5, and these allow us to calculate the required sample sizes to detect moderate mode-effects.

Given that the total sample includes two independent samples, two strategies can be used. In the first strategy, n_C is held constant and the required sample size n_M of the MM experiment is calculated, or vice versa. This strategy would be useful in the ESS because the MM experiment has been conducted additional to the main ESS data collection. In manipulating the sample size of the MM experiment, the required sample size can be calculated to detect small mode effects with a decent power. On the other hand, this strategy can also be used to detect the required sample sizes in a MM survey including a small single-mode comparative sample added to evaluate the mode effects.

The second strategy involves keeping the ratio of both n_M and n_C constant, so that they can be expressed as functions of the overall total sample size: $n_M = \lambda n$ and $n_C = (1-\lambda)n$ where $0 < \lambda < 1$. λ refers to the proportion of the total sample size which is assigned to the MM design. When λ is kept constant, the required overall total sample size n can be calculated to achieve the preferred power.

Let us illustrate both strategies of sample size calculation for all mode effects on the means of the three politics-related variables. We like to detect an effect equal to 0.05 times the range of the variables, with a power of .80 and a significance level of .05 (one-sided).

In the first strategy we fix n_C at 1294, the achieved sample size of the main ESS round 4. The calculated required sample sizes for this strategy can be found in the last-but-one column of Table 6 (formulas in appendix). Some of these n_M 's are small allowing to detect moderate mode effects with a power of .80. Other n_M 's, however, mount up to approximately 1000 which means that the MM experiment should include a rather large sample. Further, it should be noted that it is impossible to detect a selection effect of 0.05 on voter turnout with a power of .80 for any possible n_M. This results from the fact that variance introduced by the main ESS ($= a_C / n_C$) is already larger than the maximum acceptable variance of the selection effect.

Using the second strategy we fix λ at 0.214 which is the contribution of n_M to the total sample size of the ESS round 4 and the MM experiment. The results of required total sample size can be found in the last column of Table 6. These results show that a total sample size of approximately 2300 respondents allows for detecting moderate mode effects with a power of .80, except for the mode effect on voter turnout. With respect to the latter, the total sample size should be almost 6000.

[Table 6 about here.]

5. 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. This kind of evaluation is quasi impossible if only a simple MM survey dataset is available. However, we showed that the presence of data from a single-mode comparative survey allows investigating selection effects and measurement effects separately.

As an illustration, we tried to detect selection effects and measurement effects in a limited sample (with phone match) of a mixed mode survey experiment within the ESS in which respondents were offered the choice to complete the questionnaire by a web questionnaire, a telephone interview or a face-to-face interview. In comparing this MM data with the data from the main ESS round 4, which is fully conducted by face-to-face interviews, we could disentangle measurement effects from selection effects on the variables political interest, perceived complexity of politics and voter turnout in the last national election.

However, this evaluation of mode effects has some limitations which need further discussion. Two of them goes with the methods used while a third limitation relates to the characteristics of the data.

The first limitation of the method refers to the definition of the measurement effect. 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. So, these effects are not calculated on the whole sample, but only on a part of the sample. The question is whether these measurement effects can be generalized to the respondents who choose the CAPI mode in the MM survey.

The second limitation of the method might be more stringent. 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, which means that nonresponse bias should be equal in both samples. This 'comparability assumption' is probably the weakest part of the method because it is immediately contradictory with some starting points of MM designs, namely to increase coverage and nonresponse. However, we put forward some arguments supporting the use of this assumption.

Nevertheless, the next question is how much bias the comparability assumption introduces on the mode effect estimates. This can be a topic for further research. One way to answer this question might be the research on, what we call, mode-acceptance, i.e. the willingness of a respondent to participate in a particular mode, regardless whether this is his most preferred mode. With respect to the ESS example, the comparability assumption indeed involves the assumption that CATI & CAWI choosers would also accept CAPI. So, 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 non-CAT/WI choosing sample members (i.e. follow up, number of contacts, …), one can easily assume that both samples are comparable.

Last, there is also a limitation related with the data we used. 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, significant measurement effects and selection 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 ESS design, these latter two survey modes are inextricably confounded in the conclusions, it is completely impossible to measure differences between them. 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 specific to CATI and CAWI respectively, may counteract. As a result the overall effect of a combination of both modes may become small.

In summary, the presence of three survey modes in the MM data 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 a result, exact differences in measurement error or bias can be evaluated between these two modes.

References:

Agresti, A. (2002). *Categorical data analysis*. Hoboken, N.J.: Wiley.

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.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury, Calif.: Pacific Grove.

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.

Deming, W. E., & Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics, 11*, 427-444.

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.

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

Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2000). *A SAS Macro for Balancing a Weighted Sample.* Paper presented at the Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Paper 275.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41-55.

Sato, T., & Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology, 14*, 680-686.

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.

Appendix

1. Defining the variables

There are two independent samples: A MM sample (the MM experiment data) and a comparative sample (the main ESS round 4 data). We will refer to these separate samples by the subscripts M and C respectively. The total sample size n, is the sum of the MM sample size n_M and the comparative sample size n_C. These are further assumed to be constant, which means that we ignore the occurrence of systematic non-response (= comparability assumption).

We first disentangle the multinomial distributions of the variable of interest I containing J categories in both datasets to a set of Bernoulli variables. In the comparative sample we define the set of variables:

$$I_{C,j} = \begin{cases} 1 & \text{if } I = j \\ 0 & \text{otherwise} \end{cases}, \text{ for } j = 1,\ldots,J \ .$$

The $I_{C,j}$'s follow Bernoulli distributions,

$$I_{C,j} \sim b\left(\pi_{C,j}\right).$$

The sample estimates of $\pi_{C,j}$, $\hat{p}_{C,j}$, follow an asymptotic normal distribution (Agresti, 2002; Casella & Berger, 2002, p. 474) with means $\pi_{C,j}$ and covariance-matrix $\mathbf{\Psi}_C = \left(\sigma_{Cjk}\right)$ with

$$\begin{aligned} \sigma_{Cjj} &= \frac{\pi_{C,j}\left(1-\pi_{C,j}\right)}{n_C}, \\ \sigma_{Cjk} &= -\frac{\pi_{C,j}\pi_{C,k}}{n_C}. \end{aligned} \qquad (2.1)$$

In the MM sample we define the set of variables

$$I_{M,ij} = \begin{cases} 1 & \text{if } M = i \text{ and } I = j \\ 0 & \text{otherwise} \end{cases}, \text{ for } i = 1,2 \text{ and } j = 1,\ldots,J.$$

The $I_{M,ij}$'s follow Bernoulli distributions,

$$I_{M,ij} \sim b\left(\pi_{M,ij}\right),$$

whose parameter estimates $\hat{p}_{M,ij}$ follow an asymptotic normal distribution with means $\pi_{M,ij}$ and variance-matrix $\mathbf{\Psi}_M = \left(\sigma_{Mijkl}\right)$ with

$$\sigma_{Mijij} = \frac{\pi_{M,ij}\left(1-\pi_{M,ij}\right)}{n_M},$$
$$\sigma_{Mijkl} = -\frac{\pi_{M,ij}\pi_{M,kl}}{n_M}. \tag{2.2}$$

We assume that the comparative sample and the MM sample are independent, which implies that $\hat{p}_{C,k}$ is independent from $\hat{p}_{M,ij}$ for all possible i's, j's and k's, hence, their covariance is zero. We define the total variance covariance matrix as

$$\mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_C \end{bmatrix}. \tag{2.3}$$

Further we know that

$$\begin{aligned}
&P\left(I_P = j\right) = \pi_{C,j}, \\
&P\left(I_P = j \middle| M = 1\right) = \frac{\pi_{M,1j}}{\sum_k \pi_{M,1k}}, \\
&P\left(I_{WT} = j \middle| M = 2\right) = \frac{\pi_{M,2j}}{\sum_k \pi_{M,2k}}, \\
&P\left(M = 1\right) = \sum_k \pi_{M,1k} = 1 - \sum_k \pi_{M,2k}, \text{ and} \\
&P\left(M = 2\right) = \sum_k \pi_{M,2k} = 1 - \sum_k \pi_{M,1k}.
\end{aligned} \tag{2.4}$$

To end , let us define

$$\tau = \sum_k \pi_{M,1k} \ , \tag{2.5}$$

with the following first-order partial derivatives, which are useful in later derivations:

$$
\begin{aligned}
&\forall j : \frac{\partial \tau}{\partial \pi_{M,1j}} = \frac{\partial \sum_k \pi_{M,1k}}{\partial \pi_{M,1j}} = 1, \\
&\forall j : \frac{\partial \left(1-\tau\right)}{\partial \pi_{M,1j}} = \frac{\partial \left(1-\sum_k \pi_{M,1k}\right)}{\partial \pi_{M,1j}} = -1, \\
&\forall j : \frac{\partial 1/\left(1-\tau\right)}{\partial \pi_{M,1j}} = \frac{1}{\left(1-\tau\right)^2}, \text{ and} \\
&\forall j : \frac{\partial \tau\left(1-\tau\right)}{\partial \pi_{M,1j}} = 1-2\tau.
\end{aligned}
\qquad (2.6)
$$

2. Sampling distribution of the mode effects on multinomial proportions

2.1. Selection effect

Substituting (2.4) into (1.4) we get that

$$
\Sigma_j = \frac{1}{1-\tau}\left[\frac{\pi_{M,1j}}{\tau} - \pi_{C,j}\right] \qquad (2.7)
$$

Thus, the selection effect is defined by the function $s_j\left(\cdot\right)$, with first-order partial derivatives:

$$
\begin{aligned}
&\frac{\partial s_j}{\partial \pi_{M,1j}} = \frac{\tau - \tau^2}{\tau^2\left(1-\tau\right)^2} + \frac{\left(2\tau - 1\right)\pi_{M,1j}}{\tau^2\left(1-\tau\right)^2} - \frac{\tau^2 \pi_{C,j}}{\tau^2\left(1-\tau\right)^2} \\
&\frac{\partial s_j}{\partial \pi_{M,1k}} = \frac{\left(2\tau - 1\right)\pi_{M,1j}}{\tau^2\left(1-\tau\right)^2} - \frac{\tau^2 \pi_{C,j}}{\tau^2\left(1-\tau\right)^2} \quad \left(j \neq k\right) \\
&\frac{\partial s_j}{\partial \pi_{M,2j}} = 0 \\
&\frac{\partial s_j}{\partial \pi_{M,2k}} = 0 \\
&\frac{\partial s_j}{\partial \pi_{C,j}} = -\frac{1}{1-\tau} \\
&\frac{\partial s_j}{\partial \pi_{C,k}} = 0 \quad \left(j \neq k\right)
\end{aligned}
$$

Let us denote by $\mathbf{\Delta}_{Sj}$ the column vector of all these derivatives: $\left[\frac{\partial s_j}{\partial \pi_{M,11}},\ldots,\frac{\partial s_j}{\partial \pi_{M,1K}},\frac{\partial s_j}{\partial \pi_{M,21}},\ldots,\frac{\partial s_j}{\partial \pi_{M,2K}}\ ,\frac{\partial s_j}{\partial \pi_{C,1}},\ldots,\frac{\partial s_j}{\partial \pi_{C,K}}\right]$. Using the delta method restricted to the first-order Taylor series approximation (Agresti, 2002, p. 577; Casella & Berger, 2002, p. 240) we find that for $\hat{S}_j = s_j\left(\hat{p}_{M,11},\ldots,\hat{p}_{M,1J},\hat{p}_{C,j}\right) \ = \frac{1}{1-\hat{t}}\left[\frac{\hat{p}_{M,1j}}{\hat{t}} - \hat{p}_{C,j}\right]$ with $\hat{t} = \sum_k \hat{p}_{M,1k}$:

$$\frac{\hat{S}_j - \Sigma_j}{\sqrt{\sigma^2_{Sj}}} \to N\left(0,1\right), \tag{2.8}$$

where

$$\sigma^2_{Sj} = \mathbf{\Delta}'_{Sj}\mathbf{\Psi}\mathbf{\Delta}_{Sj}. \tag{2.9}$$

2.2. Measurement effect

Substituting (2.4) into (1.5) we get

$$\begin{aligned}\mathrm{M}_j &= \frac{1}{1-\tau}\left(\pi_{M,1j} + \pi_{M,2j} - \pi_{C,j}\right)\\ &= m_j\left(\pi_{M,11},\ldots,\pi_{M,1K},\pi_{M,2j},\pi_{C,j}\right)\end{aligned}.$$

The measurement effect thus is defined by function $m_j\left(\cdot\right)$ with first-order partial derivatives

$$\frac{\partial m_j}{\partial \pi_{M,1j}} = \frac{1-\tau}{(1-\tau)^2} + \frac{\pi_{M,1j}}{(1-\tau)^2} + \frac{\pi_{M,2j}}{(1-\tau)^2} - \frac{\pi_{C,j}}{(1-\tau)^2}$$

$$\frac{\partial m_j}{\partial \pi_{M,1k}} = \frac{\pi_{M,1j}}{(1-\tau)^2} + \frac{\pi_{M,2j}}{(1-\tau)^2} - \frac{\pi_{C,j}}{(1-\tau)^2}$$

$$\frac{\partial m_j}{\partial \pi_{M,2j}} = \frac{1}{1-\tau}$$

$$\frac{\partial m_j}{\partial \pi_{M,2k}} = 0$$

$$\frac{\partial m_j}{\partial \pi_{C,j}} = -\frac{1}{1-\tau}$$

$$\frac{\partial m_j}{\partial \pi_{C,k}} = 0$$

Let us denote by $\mathbf{\Delta}_{Mj}$ the column vector of these derivatives, defined analogously as $\mathbf{\Delta}_{Sj}$. Using the delta-method we find for $\hat{M}_j = m_j\left(\hat{p}_{M,11},...,\hat{p}_{M,1J},\hat{p}_{M,2j},\hat{p}_{C,j}\right)$ $= \frac{1}{1-\hat{t}}\left(\hat{p}_{M,1j} + \hat{p}_{M,2j} - \hat{p}_{C,j}\right)$:

$$\frac{\hat{M}_j - \mathrm{M}_j}{\sqrt{\sigma^2_{Mj}}} \to N(0,1) \qquad (2.10)$$

Where

$$\sigma^2_{Mj} = \mathbf{\Delta}'_{Mj}\mathbf{\Psi}\mathbf{\Delta}_{Mj}. \qquad (2.11)$$

3. Sampling distribution of the mode effects on multinomial mean

In the previous section, the calculation of mode effects on multinomial proportions was elaborated. Analogously, mode effects on the mean can be derived since the mean of a multinomial categorical distribution can be expressed in function of the proportions:

$$\mu = \sum_{j=1}^{J} j * P(I = j).$$

3.1. Selection effect

Analogue to the proportions, the selection effect on the mean can be defined as

$$\begin{aligned}\Sigma_\mu &= \left(\mu_P \middle| M=1\right)-\left(\mu_P \middle| M=2\right)\\ &=\sum_{j=1}^{J} j*P\left(I_P=j \middle| M=1\right)-\sum_{j=1}^{J} j*P\left(I_P=j \middle| M=2\right)\\ &=\sum_{j=1}^{J} j*\Sigma_j\\ &=\sum_{j=1}^{J}\frac{j}{1-\tau}\left(\frac{\pi_{M,1j}}{\tau}-\pi_{C,j}\right)\\ &=s_\mu\left(\pi_{M,11},\ldots,\pi_{M,1K},\pi_{C,1},\ldots,\pi_{C,K}\right)\end{aligned}$$

The partial derivatives are

$$\begin{aligned}Ds_\mu\left(\pi_{M,1j}\right)&=\frac{\partial s_\mu}{\partial\pi_{M,1j}}=\sum_{k=1}^{J}k*Ds_k\left(\pi_{M,1j}\right),\\ Ds_\mu\left(\pi_{M,2j}\right)&=\frac{\partial s_\mu}{\partial\pi_{M,2j}}=0,\\ Ds_\mu\left(\pi_{C,j}\right)&=\frac{\partial s_\mu}{\partial\pi_{C,j}}=-j\frac{1}{1-\tau},\end{aligned}$$

and $\mathbf{\Delta}_{S\mu}$ denotes the column vector of these derivatives. According to the delta-method the sample estimator $\hat{S}_\mu=s_\mu\left(\hat{p}_{M,11},\ldots,\hat{p}_{M,1K},\hat{p}_{C,1},\ldots,\hat{p}_{C,K}\right)$ satisfies

$$\frac{\hat{S}_\mu-\Sigma_\mu}{\sqrt{\sigma^2_{S\mu}}}\to N\left(0,1\right),$$

with

$$\sigma^2_{S\mu}=\mathbf{\Delta}'_{S\mu}\mathbf{\Psi}\mathbf{\Delta}_{S\mu}.$$

3.2. Measurement effect

The measurement effect on the mean is defined as

$$\begin{aligned}\mathrm{M}_{\mu} &= \left(\mu_{WT}\middle|M=2\right)-\left(\mu_{P}\middle|M=2\right)\\ &=\sum_{j=1}^{J} j*P\left(I_{WT}=j\middle|M=2\right)-\sum_{j=1}^{J} j*P\left(I_{P}=j\middle|M=2\right)\\ &=\sum_{j=1}^{J} j\mathrm{M}_{j}\\ &=\sum_{j=1}^{J}\frac{j}{1-\tau}\left(\pi_{M,1j}+\pi_{M,2j}-\pi_{C,i}\right)\\ &=m_{\mu}\left(\pi_{M,11},...,\pi_{M,1J},\pi_{M,21},...,\pi_{M,2J},\pi_{C,1},...,\pi_{C,J}\right).\end{aligned}$$

With partial derivatives

$$\begin{aligned} Dm_{\mu}\left(\pi_{M,1j}\right) &= \frac{\partial m_{\mu}}{\partial\pi_{M,1j}}=\sum_{k=1}^{J}k*Dm_{k}\left(\pi_{M,1j}\right),\\ Dm_{\mu}\left(\pi_{M,2j}\right) &= \frac{\partial m_{\mu}}{\partial\pi_{M,2j}}=j\frac{1}{1-\tau},\\ Dm_{\mu}\left(\pi_{C,j}\right) &= \frac{\partial m_{\mu}}{\partial\pi_{C,j}}=-j\frac{1}{1-\tau},\end{aligned}$$

where $\mathbf{\Delta}_{M\mu}$ is the column vector of these derivatives. The sample estimator of M_{μ}, $\hat{M}_{\mu}=m_{\mu}\left(\hat{p}_{M,11},...,\hat{p}_{M,1J},\hat{p}_{M,21},...,\hat{p}_{M,2J},\hat{p}_{C,1},...,\hat{p}_{C,J}\right)$ satisfies

$$\frac{\hat{M}_{\mu}-\mathrm{M}_{\mu}}{\sqrt{\sigma^{2}_{M\mu}}}\to N\left(0,1\right),$$

with

$$\sigma^{2}_{M\mu}=\mathbf{\Delta}'_{M\mu}\mathbf{\Psi}\mathbf{\Delta}_{M\mu}.$$

4. Required sample size calculations/ Power issues

The variance of the mode effects is affected by the sample sizes through the variances and the covariances of the sample proportions in $\mathbf{\Psi}$, but not by the derivatives in $\mathbf{\Delta}$. Let us define the variance-covariance matrix $\mathbf{\Psi}_{C,obs}$ and $\mathbf{\Psi}_{M,obs}$,

$$\mathbf{\Psi}_{C,obs} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & n_C \mathbf{\Psi}_C \end{bmatrix},$$

and

$$\mathbf{\Psi}_{M,obs} = \begin{bmatrix} n_M \mathbf{\Psi}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

both of dimension $3 \cdot J \times 3 \cdot J$. $\mathbf{\Psi}_{C,obs}$ and $\mathbf{\Psi}_{M,obs}$ then contain the variances and covariances of the proportion estimations for one observation in one of both samples. Using the former formulas, it can be easily shown that all the sampling variances of the mode effects are of the form

$$\sigma^2 = \frac{a_C}{n_C} + \frac{a_M}{n_M}, \tag{2.12}$$

where

$$a_C = \mathbf{\Delta}' \mathbf{\Psi}_{C,obs} \mathbf{\Delta},$$

and

$$a_M = \mathbf{\Delta}' \mathbf{\Psi}_{M,obs} \mathbf{\Delta}.$$

Given the estimation of a_C and a_M from the data, formula (2.12) can easily be used to calculate the required sample sizes by substituting it into formula (1.7).

In the case of a one sided test with significance level 0.05 (as in our example), the required $Z_{\beta,\alpha}$ to get a power of 0.80 equals

$$\begin{aligned} Z_{\beta,\alpha} &= \Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta) \\ &= \Phi^{-1}(.95) + \Phi^{-1}(.80) \\ &= 2.48, \end{aligned}$$

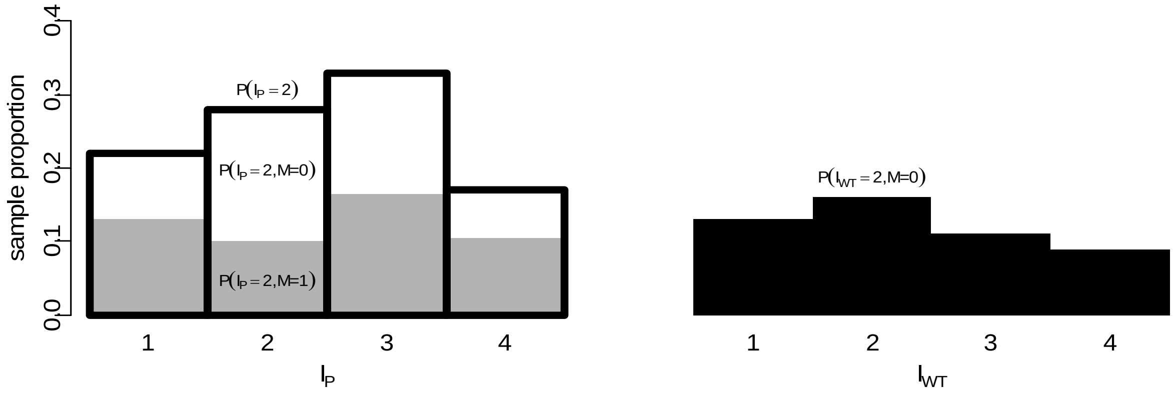
where Φ^{-1} is the inverse cumulative normal function. In the first strategy, where we fix n_C, the required n_M is calculated by

$$n_M = \frac{a_M}{\left(\frac{|effect|}{Z_{\beta,\alpha}}\right)^2 - \frac{a_C}{n_C}}.$$

In the second strategy, where λ is fixed, the total required sample size is calculated by

$$n = \frac{\lambda \cdot a_C + (1-\lambda) a_M}{(1-\lambda)\lambda\left(\frac{|effect|}{Z_{\beta,\alpha}}\right)^2}.$$

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 contoured bars in the barplot of I_P represent the proportion of the political interest scores in the ESS round 4 (i.e. $P(I_P)$). The grey areas represent the proportion of the interest scores of the respondents who answered by CAPI in the MM experiment ($P(I_P,M=1)$). The barplot of I_{WT}, represents the proportions of the respondents who answered by CAWI or CATI in the MM experiment ($P(I_{WT},M=0)$).

Table 1: Response frequencies and response rates

	ESS MM exp.	ESS round 4
CAWI	160	
CATI	88	
CAPI	104	1294
total response	352	1294
partial response	15	72
nonresponse	313	1022
noncontact	108	125
not eligible	90	161
total sample	878	2674
response rate*	44,67%	51,49%

based on sample members with matched phone number only

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

Table 2: Sample proportions

	MM exp.		
	CATI/CAWI	CAPI	ESS r4
Political Interest			
P(not at all interested)	0,084	0,033	0,067
P(hardly interested)	0,330	0,188	0,224
P(quite interested)	0,488	0,679	0,607
P(very interested)	0,098	0,100	0,101
mean	2,600	2,846	2,743
Political complexity			
P(never)	0,113	0,007	0,082
P(seldom)	0,171	0,136	0,269
P(occasionally)	0,379	0,518	0,355
P(regularly)	0,236	0,297	0,208
P(frequently)	0,102	0,042	0,085
mean	3,043	3,231	2,947
Voter turnout			
P(voted)	0,857	0,826	0,854
P(M=1)	0,255		

Table 3: Mode effects on political interest

	effect	SE(effect)	p two side	p one side	a_M	a_C
MEASUREMENT EFFECT						
P(not at all interested)	0,005	0,021	0,823	0,412	0,118	0,113
P(hardly interested)	0,093	0,037	0,012	0,006	0,368	0,313
P(quite interested)	-0,094	0,041	0,023	0,012	0,439	0,430
P(very interested)	-0,004	0,025	0,877	0,439	0,160	0,164
mean	-0,107	0,062	0,086	0,043	0,998	0,951
SELECTION EFFECT						
P(not at all interested)	-0,046	0,028	0,100	0,050	0,224	0,113
P(hardly interested)	-0,049	0,060	0,420	0,210	1,080	0,313
P(quite interested)	0,097	0,072	0,178	0,089	1,542	0,430
P(very interested)	-0,002	0,046	0,964	0,482	0,635	0,164
mean	0,139	0,098	0,154	0,077	2,797	0,951

Table 4: Mode effects on perceived political complexity

	effect	SE(effect)	p two side	p one side	a_M	a_C
MEASUREMENT EFFECT						
P(never)	0,005	0,023	0,815	0,408	0,141	0,135
P(seldom)	-0,144	0,033	0,000	0,000	0,255	0,354
P(occasionally)	0,080	0,042	0,056	0,028	0,447	0,413
P(regularly)	0,058	0,036	0,112	0,056	0,342	0,297
P(frequently)	0,002	0,024	0,945	0,473	0,142	0,141
mean	0,194	0,089	0,029	0,015	2,009	2,059
SELECTION EFFECT						
P(never)	-0,100	0,017	0,000	0,000	0,054	0,135
P(seldom)	-0,179	0,054	0,001	0,001	0,841	0,354
P(occasionally)	0,218	0,077	0,005	0,003	1,781	0,413
P(regularly)	0,118	0,070	0,090	0,045	1,479	0,297
P(frequently)	-0,058	0,032	0,070	0,035	0,288	0,141
mean	0,382	0,121	0,002	0,001	4,134	2,059

Table 5: Mode effects on voter turnout

	effect	SE(effect)	p two side	p one side	a_M	a_C
MEASUREMENT EFFECT						
P(voted)	-0,006	0,030	0,835	0,418	0,231	0,225
SELECTION EFFECT						
P(voted)	-0,037	0,058	0,523	0,262	1,016	0,225

Table 6: Required sample sizes to detect effect sizes with power=.80 and significance level=.05

Variable	effect	required effect size	n_M*	n°
pol. Intr.	meas. eff.	0,15	332	1572
	sel. eff.	0,15	644	2201
pol. Comp.	meas. eff.	0,2	419	1884
	sel.eff.	0,2	629	2302
vote	meas. eff.	0,05	998	3331
	sel. eff.	0,05	N.A.	5801

*: *keeping n_C constant*

°: *keeping λ constant*

N.A.: impossible to estimate