Public Opinion Quarterly, Vol. 74, No. 5, 2010, pp. 1027–1045

A METHOD FOR EVALUATING MODE EFFECTS IN MIXED-MODE SURVEYS

JORRE VANNIEUWENHUYZE*
GEERT LOOSVELDT
GEERT MOLENBERGHS

Abstract Survey designs in which data from different groups of respondents are collected by different survey modes have become increasingly popular. However, such mixed-mode (MM) designs lead to a confounding of selection effects and measurement effects (measurement error) caused by mode differences. Consequently, MM data have poor quality. Nevertheless, comparing MM data with data from a comparable single-mode survey allows researchers to measure selection effects and measurement effects separately. The authors develop a method to evaluate mode effects and illustrate this method with 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.

Introduction

Increasingly, data are gathered by mixing different survey modes in one design (Dillman, Smyth, and Christian 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 an MM data collection can be

JORRE VANNIEUWENHUYZE is a Ph.D. candidate at the Centre for Sociological Research, Katholieke Universiteit Leuven, Belgium. GEERT LOOSVELDT is Professor of Survey Methodology at the Centre for Sociological Research, Katholieke Universiteit Leuven, Belgium. GEERT MOLENBERGHS is Professor of Biostatistics at I-BioStat, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium.
A previous version of this article was presented at the International Total Survey Error Workshop (ITSEW), in Stowe, Vermont, USA, in June 2010. The research is part of the European Social Survey Infrastructure Preparatory Phase Project [ESSPrep, Contract Number 212311]. *Address correspondence to Jorre Vannieuwenhuyze, OE Centrum voor Sociologisch Onderzoek, Parkstraat 45, Bus 3601, 3000 Leuven, Belgium; e-mail: jorre.vannieuwenhuyze@soc.kuleuven.be.

doi: 10.1093/poq/nfq059

advantageous in several ways (de Leeuw 2005; Dillman, Smyth, and Christian 2009). First, it can help reduce coverage error because several modes are available to contact different groups of hard-to-reach respondents. Second, an MM data collection can help lower non-response and non-response bias in order to reduce the Total Survey Error (TSE) because every respondent can choose his or her mode of preference among several modes. Third, MM data collection can help reduce costs because a substantial part of the sample will be surveyed by an inexpensive mode.

However, notwithstanding their advantages, MM designs do not automatically lead to higher data quality or smaller TSE (Voogt and Saris 2005). MM designs may lower non-response 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 selection effects and measurement effects (measurement error).

Selection effects occur when different types of respondents choose different modes to complete the survey. As such, they are forms of non-response 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 an MM design valuable. Indeed, because of selection effects, some respondents may accept participation when they would not (non-response) or could not (non-coverage) in a single-mode survey (Biemer 2001; Day, Dunt, and Day 1995; de Leeuw and Van Der Zouwen 1988; Dillman et al. 2009; Voogt and 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, so that one person would give different answers in different modes (Bowling 2005; Voogt and Saris 2005; Weisberg 2005). 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, and acquiescence (Bowling 2005; Brick and Lepkowski 2008; de Leeuw 1992; de Leeuw 2005; Dillman 1991; Dillman, Smyth, and Christian 2009; Schwarz et al. 1991).

In order to evaluate the TSE introduced by an 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, and education level) to disentangle both mode effects (de Leeuw 2005; Jäckle, Roberts, and Lynn 2010). This method assumes that the matching variables are closely related with the variables of interest, but this

assumption can hardly be supported. 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 develop a method to disentangle measurement effects from selection effects on the proportions and the mean of a multinomial variable by comparing an MM dataset with a comparable single-mode dataset. This method will be introduced in the next section. Subsequently, the methods will be illustrated with mixed-mode data from the European Social Survey by calculating the mode effects on the parameters of three politics-related variables.

A Method to Disentangle Mode Effects in a Mixed-mode Dataset Using Comparable Single-mode Data

Let us assume we have a mixed-mode (MM) dataset of size n_m where some respondents responded by mode A and others responded by mode B. Let us further assume that we also have a single-mode dataset where all respondents responded by mode A. We will call this dataset the comparative dataset because these data will be compared with the data from the MM sample. Let n_c denote the sample size of this comparative sample, and $n = n_m + n_c$ the total sample size.

Further, we denote by Y the multinomial variable of interest with J categories. Two versions of this variable can be distinguished, namely Y_a and Y_b, where Y_a refers to the values of Y when this variable is observed by mode A, while Y_b refers to the same variable though observed by mode B. We assume that each population member takes values on both these variables, and that these values are not necessarily the same for each person. Considering the outcome of different survey modes as different variables allows us to evaluate measurement effects merely by comparing Y_a and Y_b. Of course, given the survey design, either Y_a or Y_b is observed for each respondent and this problem should be circumvented. Both Y_a and Y_b follow a multinomial distribution with parameter vector $\boldsymbol{\pi}_{\mathbf{m}} = (\pi_{m1}, \ldots, \pi_{mJ})$, where $m = a$ or b, respectively.

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. Thus, M is a binary variable with values a (mode A) or b (mode B) following a Bernoulli distribution.

REPRESENTATIVITY ASSUMPTION

As already noted, our method to evaluate mode effects involves comparing the MM sample with the comparative sample. However, in doing so, we implicitly assume that the realized samples (MM and comparative) represent the same population. Put differently, we assume that differences in the distribution of

the unbiased version of the variable(s) of interest are only caused by sampling error (or purely random non-response and coverage error). We call this assumption the "*representativity assumption*." Differences in systematic coverage error can usually be evaluated easily by comparing how the sampling frame was set up in both survey designs. Unfortunately, differences in systematic non-response error, in contrast, can generally not be evaluated directly. Nevertheless, two arguments can be put forward to substantiate this assumption.

First, if both samples contain a comparable set of respondents, all respondents of the MM sample, responding by either mode A or B, would also accept participation in a single-mode survey completely conducted by mode A. In some situations, this assumption is reasonable given the modes used, as our example in Section 3 illustrates. As a consequence, we expect the difference between the response rates to be zero, and this difference can be tested statistically. A nonsignificant difference in response rates is an argument enforcing the representativity assumption. Still, a comparison of response rates as an argument for the representativity assumption is not decisive because both samples may have attracted different respondents by putting in different effort to reach certain types of respondents.

A second argument for the representativity assumption involves a comparison of the composition of both datasets on a set of "mode-insensitive" socio-demographical variables. If both samples turn out to have a comparable composition, this can be used as an additional argument in favor of the representativity 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.

DEFINING THE MODE EFFECTS

We can now define the selection effect on the proportion parameter of Category j as the difference between this proportion measured by the same mode, but observed on the two different groups of respondents, namely those who would answer by mode A and those who would answer by mode B in the MM sample. If we choose mode A as the standard mode, the selection effect $S_a(\pi_j)$ on proportion π_j of category j can be defined as follows:

$$S_a(\pi_j) = P(Y_a = j|M = a) - P(Y_a = j|M = b). \tag{1.1}$$

Next, we can define the measurement effect $M(\pi_j)$ on the proportion parameter π_j of category j as the difference between the measures of this proportion obtained by the two different modes, though observed on the same group of respondents. If this group of respondents is the respondents who would choose mode B (i.e., $M = b$), the measurement effect is equal to

$$M_b(\pi_j) = P(Y_b = j|M = b) - P(Y_a = j|M = b). \tag{1.2}$$

In both these definitions, $P(Y_a = j|M = a)$ and $P(Y_b = j|M = b)$ can simply be estimated with the MM data. $P(Y_a = j|M = b)$, however, is never observed directly because Y_a is not measured for the respondents who chose to answer by mode B. Nonetheless, we can use the law of total probability to prove that

$$P(Y_a = j|M = b) = P(Y_a = j)\frac{1}{P(M = b)} - P(Y_a = j|M = a)\frac{P(M = a)}{P(M = b)}. \tag{1.3}$$

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

$$\mathrm{S}_a(\boldsymbol{\pi_j}) = \frac{1}{P(M = b)}[P(Y_a = j|M = a) - P(Y_a = j)] \tag{1.4}$$

and

$$\begin{aligned}\mathrm{M}_b(\boldsymbol{\pi_j}) = {} & P(Y_b = j|M = b) - P(Y_a = j)\frac{1}{P(M = b)} \\ & + P(Y_a = j|M = a)\frac{P(M = a)}{P(M = b)}.\end{aligned} \tag{1.5}$$

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

- $P(Y_a)$ from the comparative dataset, which is a sample completely surveyed by mode A.
- $P(Y_a|M = a)$ from the MM data, more specifically from the respondents who responded by mode A.
- $P(Y_b|M = b)$ from the MM data as well, but now from the respondents who responded by mode B.
- $P(M = a)$ and $P(M = b)$ from the whole MM dataset.

Sometimes variable Y is a scale variable where the categories can be ordered and the difference between every two adjacent categories can be assumed to be equal. In that situation, we can also define the mode effects on the mean, because the mean can be expressed as a function of the proportions:

$$\mu_m = \sum_{j=1}^{J} j\pi_{mj} \text{ for } m = a \text{ or b.} \tag{1.6}$$

It can be shown that the selection effects on the mean equals

$$\begin{aligned} S_a(\mu) &= (\mu_a|M = a) - (\mu_a|M = b) \\ &= \sum_{j=1}^{J} j\Sigma(\pi_j), \end{aligned} \tag{1.7}$$

and the measurement effect on the mean is

$$\begin{aligned} M_b(\mu) &= (\mu_b|M = b) - (\mu_a|M = b) \\ &= \sum_{j=1}^{J} jM(\pi_j). \end{aligned} \tag{1.8}$$

All mode effects, as defined in (1.4), (1.5), (1.7), and (1.8), are transformations of proportion parameters. All these proportions can be estimated from the sample data, and their sampling distribution is known to be asymptotically normal (Agresti 2002; Casella and Berger 2002). The sampling variances and covariances of these proportion estimates can also be calculated easily. Given these properties, the Delta method restricted to the first-order Taylor series approximation (Agresti 2002; Casella and Berger 2002) proves that the selection and measurement effects are asymptotically normal as well, and provides approximations of their sampling variances. For a detailed overview of these calculations, we refer to the technical note of Vannieuwenhuyze and Molenberghs (2010).

REQUIRED SAMPLE SIZE CALCULATIONS/POWER ISSUES

An additional question that should be asked when evaluating the mode effects is whether the total sample size n is sufficiently large to detect small to medium mode-effect sizes. Let θ denote the size of the mode effect we want to detect with a minimal power β, given that we use a significance level α. θ corresponds to a z-value

$$z = \theta/\sqrt{\sigma^2}, \tag{1.9}$$

where σ^2 is the sampling variance of the mode-effect estimate. The absolute value of this z-value should at least be equal to $\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)$, where Φ^{-1} is the inverse cumulative normal function. $\Phi^{-1}(\alpha)$ corresponds to the minimal z-value to detect a significant effect with significance level α. $\Phi^{-1}(\beta)$ is the difference between $\Phi^{-1}(\alpha)$ and the required z-value of θ so that θ is detected with a minimal power β. For example, if we like to detect a mode effect with a power of 0.80 while it is evaluated with a one-sided test with significance level α of 0.95, the z-value corresponding with θ should be

$$|z| \geq \Phi^{-1}(0.95) + \Phi^{-1}(0.80) = 1.64 + 0.84 = 2.48. \tag{1.10}$$

Further, using the properties of the Delta method (Agresti 2002; Casella and Berger 2002), it can be shown that all the sampling variances of the mode effects are of the form

$$\sigma^2 = \frac{a_c}{n_c} + \frac{a_m}{n_m}. \tag{1.11}$$

In this equation, a_c/n_c and a_m/n_m represent the contribution of, respectively, the comparative and the mixed-mode sample to the sampling variance of the mode-effect estimates. a_c and a_m can be calculated in analogy with the Delta method but using a covariance matrix for the sample proportions, which is not corrected for the sample sizes of both samples. As a result, these statistics do not depend on these sample sizes. The exact formulas of a_c and a_m can be found in the technical note of Vannieuwenhuyze and Molenberghs (2010) as well.

Given the estimation of a_c and a_m from the data, implementing (1.11) into (1.9) allows calculating the minimal required sample sizes to achieve a decent power given the critical significance level:

$$\frac{\theta^2}{\frac{a_c}{n_c} + \frac{a_m}{n_m}} \geq \left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\right]^2. \tag{1.12}$$

Because the total sample includes two independent samples, two strategies can be used. In the first strategy, the sample size of the mixed-mode sample, n_m, or the comparative sample, n_c, is held constant, and the required sample size of the other sample is calculated by rearranging the terms in (1.12). For a fixed sample size of the comparative group, the minimal sample size of the mixed-mode sample becomes

$$n_m \geq a_m \left(\frac{\theta^2}{\left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\right]^2} - \frac{a_c}{n_c} \right)^{-1}. \tag{1.13}$$

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 total sample size n to achieve the preferred power can be calculated by

$$n \geq (\lambda a_c + (1 - \lambda)a_m) \left(\frac{\theta^2}{\left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\right]^2} \cdot (1 - \lambda)\lambda \right)^{-1}. \tag{1.14}$$

An Illustration with ESS Data

THE ESS AND THE 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 behavior patterns of its diverse populations. It contains topics like 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 most recent 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 (CAPI) so far. Because of the costs of FTF interviews, declining funds, declining response rates, changing coverage issues, and the resistance from certain countries without a tradition of conducting FTF interviews, an MM experiment was set up in the Netherlands parallel to the fourth round (Eva et al. 2010). 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 illustration, we will use the 2,674 sample members of the main Dutch ESS data who could be matched to a telephone number in the sampling list. These respondents were reached by at most ten interviewer contact attempts at home. In the MM experiment, a sample of 878 persons with a matched phone number was drawn from the 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 telephone interview (CATI), or a face-to-face personal interview at home (CAPI)—from the very first contact.[1] Sample members without a matched telephone were also included in both the main ESS and the MM experiment, but almost all of the experiment respondents responded by an FTF interview as well. Consequently, this group is hardly useful for evaluating mode effects.

Both samples contain a simple random sample 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.

The MM experiment started with a telephone contact (first telephone screening) including 14 call attempts. If a person was willing to participate in 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 first screening were subject to a second telephone screening, which was performed analogously to the first screening.

The follow-up of non-response depended on the mode each respondent chose in the telephone screenings. First, the respondents who chose to complete the Web questionnaire were recontacted and reminded at most 14 times by telephone

1. The MM experiment also contains a sequential design in which modes were offered sequentially (first Web, then telephone, then FTF) instead of simultaneously. However, we restrict our analyses to the concurrent MM data.

if needed. If a respondent refused to complete the Web questionnaire, a telephone or FTF interview was still offered. Nonetheless, these non-respondents were not automatically recontacted at their homes by an interviewer.

Second, the sample members who chose a telephone interview were either interviewed immediately during the telephone screening or an appointment was made for a callback. Although these sample members were allowed to change their mind and ask for a Web questionnaire or an FTF interview, only one switched to a Web survey. Non-response could occur if there was no contact at an appointment. These non-respondents were approached FTF for a personal interview in a follow-up phase after the telephone screening phase.

Finally, the respondents who chose an FTF interview were visited by an interviewer at home. Non-contacts or non-response were not followed up in any other survey mode.

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

Response frequencies of both datasets can be found in Table 1. For convenience, respondents with partially incomplete answers to the variables described in the next section were left out for the further analyses. Both the main ESS data and the MM experimental 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

Table 1. Response Frequencies and Response Rates

	ESS MM exp	ESS round 4
CAW1	160	
CAT1	88	
CAP1	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%

NOTE.—Based on sample members with matched phone number only, * = total response/(total samle - not eligible)

2. See http://www.cbs.nl.

post-stratification weights were calculated using iterative proportional fitting or raking procedures (Deming and Stephan 1940; Izrael, Hoaglin, and Battaglia 2000).

To conclude, 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, our method only allows for evaluating differences between CAPI (mode A) on the one hand and a combination of CATI and CAWI (mode B) on the other. The latter two modes cannot be compared to each other without additional assumptions. As a consequence, mode A corresponds in a certain way to a single-mode CAPI survey, while mode B corresponds to a concurrent mixed-mode CATI-CAWI survey. The measurement effects then represent the differences between the parameter estimates of both these surveys. The selection effects, on the other hand, represent the differences between the respondents who chose CAPI and those who chose CATI or CAWI in a three-mode design, but on the parameter estimates that would be obtained with a two-mode CATI-CAWI survey for all these sample members (i.e., mode B). This specific problem would not have happened if the MM experiment contained only two modes (CAPI and any other mode).

CHECKING THE REPRESENTATIVITY ASSUMPTION

Since both samples are drawn from the same sampling frame, there can be no difference in systematic coverage error. Further, it is well known and generally observed that CAPI 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, it makes sense to theoretically assume that those who chose the CAWI and CATI in the MM experiment would also agree to participate in an FTF survey when they were sampled for the main ESS Round 4 data collection.

However, the response rate of the ESS MM experiment is, remarkably, significantly lower than the response rate of the main ESS survey (±7%; see Table 1). This difference is probably caused by differences between the two surveys in efforts made to reach all sample members. Sample members of the MM experiment who chose to participate by CAWI but did not respond were not followed up by CAPI. This inaccuracy in sample design might explain the difference in response rates.

On the other hand, 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 did not show any significant difference (tables not shown here). This can be used as an argument enforcing the representativity assumption. Nevertheless, we corrected for the small remaining differences using normalized propensity-score

weights derived from the complete set of variables mentioned above (Rosenbaum and Rubin 1983; Sato and Matsuyama 2003). As a consequence, both datasets are comparable on these socio-demographic characteristics.

VARIABLES

In this illustration, we will separately analyze 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 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, 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 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 show card with all five substantial answer categories. In the CATI mode, the question and answers were read to the respondent analogous to CAPI, but no show cards were used. In the CAWI mode, the questions were shown using the same wording and order of answer categories. If, however, the respondent tried to skip a question, a "don't know" answer appeared at the bottom of the answer list. The respondent was obliged to select one answer.

All 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 and Van Kempen 2002). It has been argued that measurement effects are strongest on questions about such socially desirable behaviors (Brick and Lepowski 2008; Schwarz et al. 1991; Voogt and Saris 2005; Weisberg 2005). Because of the present interaction between interviewer and respondent, respondents act according to social norms and give cultural acceptable answers in an interview survey. As a consequence, we expect that people tend to overreport 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; Dillman 2005; Dillman et al. 2009; Voogt and Saris 2005; Weisberg 2005). Perceived complexity of politics and voter turnout is generally highly correlated with political interest (e.g., in ESS Round 4, the correlation between interest and perceived complexity is -0.433, $p < 0.001$; the difference in interest is 0.673 between voters and nonvoters in ESS Round 4, $p < 0.001$). Highly interested people generally evaluate politics as less complex, and voters are usually more

interested in politics. So, we expect measurement effects on these variables as well.

Second, Voogt and Van Kempen (2002) also argue that non-respondents are usually less interested in politics. Because the CAPI group of the MM experiment contains a considerable group of non-respondents in the first phase of the survey, we can expect selection effects on all three variables as well. We expect that those who chose CAPI in the MM experiment are less interested in politics, perceive politics as more complicated, and are less likely to have voted in the last election.

Results

Table 2 summarizes the observed sample proportions and means, which are used to calculate the mode-effect 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 would fall between the means of the two MM groups, provided that the representativity assumption holds. Indeed, the representativity assumption means that the ESS and the mixed-mode sample, which is the combination of the two MM groups, present the same population. The data, however, show a different trend. The mean political complexity in the main ESS is lower than in both MM groups, which might be explained by mode effects.

Table 2. Sample Proportions

	MM exp.		
	CATI/CAWI	CAPI	ESS r4
Political Internet			
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(occasinally)	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		

POLITICAL INTEREST

Table 3 shows 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 answer "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 of the category "quite interested" is negative, the opposite conclusion can be made. 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. Thus, respondents may report a higher interest in politics in front of an interviewer because this probably is socially desirable behavior (Voogt and Van Kempen 2002).

If the two-sided *p*-values of the selection effects are considered, none of the selection effects seem to be significant. We expected that the CAPI choosers in the MM design were less interested in politics because this group contains more non-respondents in the first phase of the survey (Voogt and Van Kempen 2002). This means that the selection effect on the mean should be negative, but, as Table 3 shows, this expectation is not met. Consequently, we cannot 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. 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

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 select 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 meets 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.

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,

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

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 conducted totally by CAPI. Analogously, a difference in voting behavior between CAPI choosers and CATI/CAWI choosers is not confirmed either.

SAMPLE SIZE CALCULATION

Let us now illustrate how to calculate the required sample sizes in the ESS example to detect small to moderate values of the mode effects with a specific power. For this article, we restrict our example to the means of the three politics-related variables. Let us assume that we like to detect a small mode effect equal to 0.05 times the range of the variables, with a power of .80 and a significance level of .95 (one-sided). The sample estimates of a_c and a_m can be found in Table 3, 4, and 5.

Using the first strategy, we fix the sample size of the comparative group and manipulate the sample size of the MM experiment. This would be useful in the ESS because the MM experiment has been conducted in addition to the main

Table 6. Required Sample Sizes to Detect Moderate Mode Effects with Power = .80 and Significance Level = .05

Variable	Effect	Minimal effect	n_m*	n°
Minimal Sample Sizes to Detect Moderate Mode Effect:				
pol. Intr.	meas. eff.	0.15	332	1572
	sel. eff.	0.15	644	2201
pol. Comp.	meas. eff.	0.20	419	1884
	sel. eff.	0.20	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.

ESS data collection. We fix n_c at 1,294, which is the achieved sample size of the main ESS Round 4. The calculated required sample sizes n_m for this strategy can be found in the next-to-last column of Table 6. The realized sample in the mixed-mode experiment comprises 352 respondents. As the results show, this sample size was only sufficient to detect small measurement effects on the variable political interest. Other small mode effects on the means of the three variables of interest would not be detected with such a small sample size in the experiment. Some n_m's even mount up to approximately 1,000, which means that the MM experiment should include a rather large sample to be able to detect a small mode effect. 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 the variance introduced by the main ESS $(=a_c/n_c)$ is already larger than the maximum acceptable variance of the selection effect.

In the second strategy, we fix λ at 0.214, which is the contribution of n_m to the total sample size of ESS Round 4 and the MM experiment. The results of the required total sample size can be found in the last column of Table 6. These results show that a total sample size of approximately 2,300 respondents allows for detecting small 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 6,000. The actual total sample size, however, is only 1,646, which means that the realized ESS and mixed-mode experiment sample can only detect a significant small measurement effect on the variable political interest.

The sample size calculations in this section lead to the conclusion that the realized sample sizes of the ESS, the mixed-mode sample, or both are mostly too small to detect small mode effects except for the measurement effect on political interest. With a sample size of 650 instead of 352 in the mixed-mode experiment, for example, small mode effects on political interest and perceived complexity could have been detected. With respect to voter turnout, however, the sample sizes should be unreasonably large to be able to detect small mode effects.

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 an MM survey context. This kind of evaluation is quasi-impossible if only a simple MM survey dataset is available, but we showed that the presence of data from a single-mode comparative survey allows researchers to investigate selection effects and measurement effects separately. However, this evaluation of mode effects relies on some assumptions that need further discussion.

The first and probably most stringent assumption is the representativity assumption, which has already been discussed. This assumption means that *systematic* coverage and non-response error should be equal in both the MM

sample and the comparative sample. The more this assumption is violated, the more the mode-effect estimates would probably be biased. The magnitude of this bias depends on the correlation between the variable of interest and what we call the survey acceptance patterns of the sample members.

By survey acceptance patterns, we refer to the willingness of a respondent to participate in both the mixed-mode survey and the comparative survey. The larger the group of sample members who would participate in only one of the survey designs, the less both samples probably represent the same population and the more bias the method can introduce on the mode-effect estimates. The magnitude of the bias then depends on the extent to which this group of sample members differs from the sample members who would participate in both surveys. Put differently, the larger the correlation between the survey acceptance pattern and the variable of interest, the larger the bias on the mode-effect estimates.

In short, the bias on the mode-effect estimates thus depends on the survey acceptance patterns and the correlation with the variable of interest. Future research may include a sensitivity analysis to the robustness of the mode-effect estimates to fluctuations in these patterns and the correlations.

Second, our method also implies that measurement error and bias for mode A is equal in both the MM sample and the comparative sample. The particular survey design of both samples might, however, enhance differences in measurement. Nevertheless, we expect this assumption to be less stringent compared to the representativity assumption.

Further, our method has two limitations as well. The first limitation of the method refers to the definition of the measurement effect. We calculated the measurement effect using the difference between the statistics obtained in mode A and mode B, respectively, but only for the respondents who chose mode B in the MM design. Thus, these effects are calculated for only part of the sample. The question is whether these measurement effects can be generalized to the respondents who chose mode A in the MM survey.

Second, the method we offered works fine if there is only one comparative dataset and the MM data are gathered by only two modes. When the MM sample includes more than two modes (say modes A, B, and C), as in the ESS data, additional comparative samples and assumptions are required. Otherwise, modes B and C are completely confounded in the conclusions. Such an additional sample should include two of the three modes, for example A and B, so that mode effects between A and B can be estimated exactly. However, in order to estimate the exact mode effects, the researcher must assume in that situation that the respondents of modes B and C in the triple-mode sample are comparable to the mode B-choosers in the double-mode sample. This assumption might obstruct the validity of mode-effect estimates.

To conclude, we would like to make a suggestion for future surveys. Our method is applicable as soon as a mixed-mode and a comparative sample are available, with strong signs that the comparability assumption is valid or

only slightly violated. This means that the MM data may be gathered not only by a concurrent design, as in the ESS, but also by a sequential design where the modes are offered one after the other to the sample members. Such a sequential design can start with an inexpensive mode B to decrease costs as much as possible, and then a follow-up can be organized in mode A to reduce non-response as much as possible. Parallel to this MM data collection, a small comparative sample can be drawn and surveyed completely using mode A. Such an extended MM design allows for evaluating the mode effects, even though costs are reduced and non-response is probably reduced as well. An additional advantage of such a design is that the implementation of mode B can be organized with a primary focus on reducing measurement error while non-response is only a secondary concern. The implementation of mode A, in contrast, should focus on non-response reduction while measurement error is of secondary concern. To guarantee the validity of the representativity assumption, a considerable time gap between the initial and follow-up phases in the mixed-mode sample may help counter the influence of a refusal in the initial phase on participation in the follow-up.

References

Agresti, Alan. 2002. *Categorical Data Analysis*. Hoboken, NJ: Wiley.

Aquilino, William S. 1994. "Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment." *Public Opinion Quarterly* 58(2):210–40.

Biemer, Paul 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, Ann. 2005. "Mode of Questionnaire Administration Can Have Serious Effects on Data Quality." *Journal of Public Health* 27(3):281–91.

Brick, J. Michael, and James M. Lepowski. 2008. "Multiple Mode and Frame Telephone Surveys." In *Advances in Telephone Survey Methodology*, eds. James Lepowski, Lilli Japec, Clyde Tucker, Paul Lavrakas, Michael Brick, Michael Link, Edith De Leeuw, and Roberta Sangster. Hoboken, NJ: Wiley & Sons.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Duxbury, CA: Pacific Grove.

Day, Neil Atherton, David R. Dunt, and Susan Day. 1995. "Maximizing Response to Surveys in Health Program Evaluation at Minimum Cost Using Multiple Methods." *Evaluation Review* 19(4):436–50.

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

———. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21(2):233–55.

de Leeuw, Edith D., and Johannes Van Der Zouwen. 1988. "Data Quality in Telephone and Face-to-face Surveys: A Comparative Analysis." In *Telephone Survey Methodology,* eds. Robert Groves, Paul Biemer, Lars Lyberg, James Massey, William Nicholls II, and Joseph Waksberg. New York: Wiley-Interscience.

Deming, William Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *Annals of Mathematical Statistics* 11:427–44.

Dillman, Don A. 1991. "The Design and Administration of Mail Surveys." *Annual Review of Sociology* 17:225–49.

———. 2005. "Survey Mode as a Source of Instability in Responses across Surveys." *Field Methods* 17(1):30–52.

Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L. Messer. 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, Don A., Jolene D. Smyth, and Leah Melani Christian. 2009. *Internet, Mail, and Mixed-mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.

European Social Survey. 2008. " Round 4 Data." Data file edition 1.0: Norwegian Social Science Data Services, Norway, data archive and distributor of ESS data.

Eva, Gillian, Geert Loosveldt, Peter Lynn, Peter Martin, Melanie Revilla, Willem Saris, and Jorre Vannieuwenhuyze. 2010. "ESS Prep WP6 - Mixed Mode Experiment. Delivarable 21 final mode report." London: City University.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Izrael, David, David C. Hoaglin, and Michael P. Battaglia. 2000. "A SAS Macro for Balancing a Weighted Sample" Proceedings of the Twenty-fifth Annual SAS Users Group International Conference, Paper 275.

Jäckle, Anette, Caroline Roberts, and Peter Lynn. 2010. "Assessing the Effect of Data Collection Mode on Measurement." *International Statistical Review* 78:3–20.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Sato, Tosiya, and Yutaka Matsuyama. 2003. "Marginal Structural Models as a Tool for Standardization." *Epidemiology* 14:680–86.

Schwarz, Norbert., Fritz Strack, Hans-J. Hippler, and George Bishop. 1991. "The Impact of Administration Mode on Response Effects in Survey Measurement." *Applied Cognitive Psychology* 5(3):193–212.

Vannieuwenhuyze, Jorre, and Geert Molenberghs. 2010. "An SAS Macro to Disentangle Mode Effects on Proportions and the Mean of a Categorical Variable in an Extended Mixed-mode Dataset." Accessed July 22, 2010. http://perswww.kuleuven.be/jorre_vannieuwenhuyze.

Voogt, Robert J. J., and Willem E. Saris. 2005. "Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21(3):367–87.

Voogt, Robert J. J., and Hetty Van Kempen. 2002. "Nonresponse Bias and Stimulus Effects in the Dutch National Election Study." *Quality & Quantity* 36(4):325–45.

Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago.