Estimation under Mode Effects and Proxy Surveys, Accounting for Non-ignorable Nonresponse

Danny Pfeffermann[1,2,3] and Arie Preminger[1]

ABSTACT

We propose a new, model-based methodology to address two major problems in survey sampling: The first problem is known as mode effects, under which responses of sampled units possibly depend on the mode of response, whether by internet, telephone, personal interview, etc. The second problem is of proxy surveys, whereby sampled units respond not only about themselves but also for other sampled. For example, in many familiar household surveys, one member of the household provides information for all other members, possibly with measurement effects. Ignoring the existence of mode effects and/or possible measurement effects in proxy surveys could result in possible bias in point estimators and subsequent inference. Our approach accounts also for nonignorable nonresponse. We illustrate the proposed methodology by use of simulation experiments and real sample data, with known true population values.

Keywords: EM algorithm; measurement effects; NMAR nonresponse; probability and nonprobability sampling, selection effects.

1- Central Bureau of Statistics, Israel.

2- Department of Statistics, Hebrew University of Jerusalem, Israel.

3- Southampton Statistical Sciences Research Institute, University of Southampton, UK.

0. PREFACE

I felt very happy and privileged when invited to submit a paper for the special issue of Sankhya A, celebrating the 100th birth anniversary of Prof. C. R. Rao. Professor Rao contributed, indirectly, a great deal to my research. In 1993 I published an article in the *International Statistical Review* entitled, "The Role of Sampling Weights when Modeling Survey Data", (Pfeffermann, 1993). While working on this article, I came across a short discussion made by the late Professor Steve Fienberg in 1989, stating that "the one exception in which the use of weights may be appropriate is outcome-based sampling, where the sampling plan may be informative for the model of interest." Professor Fienberg referred to an earlier article by Patil and Rao (1978), which shows how the sampling weights (inverse of the sample inclusion probabilities) feature in the distribution of the sample data in such cases, and how the sample distribution differs from the corresponding population distribution. One of the examples in that article was probability proportional to size (PPS) sampling. This whole area was new to me at the time, but I got really attracted to it, and since then I published many articles with colleagues on the relationship between the population distribution, the sample distribution and the non-sample distribution, and how the latter two distributions can be used for inference about the population distribution and for imputation of missing data. See Pfeffermann (2017) for a unified theory with applications to informative sampling and nonignorable nonresponse, small area estimation, observational studies, web panels and more. The present article is another extension of this general theory.

Professor Rao contributed more directly to my work in 2016, when inviting me to co-edit with him the 29th Handbook of Statistics on *Survey Samples*. This turned out to be a fascinating experience and ended up with a two-volume handbook, containing 41 chapters spread over 1300 pages.

I am very grateful to Professor Rao for his indirect and direct contributions to my career, and I wish him many more years of happy and productive life, with good health.

1. INTRODUCTION

In modern sample surveys, sampled units often have the choice of how to respond, whether by telephone, personal interview, mail, fax, or via the Internet. Such surveys are nowadays very popular in many countries, called *mixed-mode surveys*. See, e.g., de Leeuw (2018) for a recent comprehensive review. Sometimes, the different modes of response are offered sequentially. For example, when starting the survey, all the sampled units are encouraged to respond via the internet. Those who do not respond within a certain time period are approached by telephone and finally, those who couldn't be contacted or refused to respond via the telephone, are approached for a personal interview.

The term *mode-effect* encompasses two confounded effects: *selection effect* - the effect of differences between characteristics of respondents preferring to respond with different modes and consequently, possible differences in the values of reported study variables of interest, and *measurement effect* - the effect of potentially responding differently by the same person, depending on the mode of response. The motivation behind the use of mixed mode surveys is to possibly increase the response rates and reduce measurement effects, by letting each person to reply by his preferred mode. Clearly, some modes are cheaper and simpler than other, notably, the use of the internet. The literature contains many examples illustrating that different modes of data collection can affect the responses. See also Table 4 in Section 7.1 of the present paper.

If all sampled units respond correctly by their preferred mode, no bias occurs and the use of a mixed-mode survey benefits from the advantages listed above. However, in practice, no responses are obtained from some of the sampled units, with the rates of nonresponse steadily increasing in recent years all over the world. In this case, the use of mixed-mode surveys may introduce large bias in sample estimators, if not accounted for properly. The situation is even worse in the case of measurement effects. It is often recommended to reduce the measurement effects by a careful questionnaire design across the modes, see, e.g., Dillman and Christian (2003) and de Leeuw et al.(2018), but in the present article we assume a given sample with given responses. Notably, the two effects are confounded, and several studies in the literature attempt to disentangle them, see, e.g., de Leeuw (2005), Hox et al. (2017) and Vannieuwenhuyze et al. (2010, 2014).

In order to reduce the total mode effect, it is common to first determine whether the survey estimates produced from the different modes are indeed different and if they are, to infer which mode is the best in the sense of producing the smallest bias for the variable of interest. The selected mode is then used as a benchmark for correcting the other modes. Vannieuwenhuyze, et al. (2014) assume the existence of a mode under which no bias occurs and develop bias corrected estimators by applying the observational study theory of Rosenbaum and Rubin (1983, 1984). In another approach, mode effects are conceptualized as a missing-data problem. Here again, one of the models is assumed to yield correct measurements and is used to impute values for the other modes. For example, Park et al. (2016) consider the case of two modes, use one of them as a benchmark and assume a linear relationship between the observations obtained under the two modes.

The mode comparisons are often based on heuristic arguments. For example, for questions on sensitive topics such as drug use and alcohol consumption, it is sometimes assumed that the mode which provides the highest prevalence of the illicit behavior produces the smallest bias, since the tendency of respondents would be to underreport such behavior, (Tourangeau and Yan, 2007). An obvious shortcoming of this approach is the underlying assumption that there consists a tendency to underreport sensitive questions. (Turner et al. 1998).

An alternative approach to assess mode effects is to compare the estimates obtained from the different modes with known external data, which is assumed to be more accurate. For example, in an income study in Denmark, Kormendi (1988) estimated the bias obtained from the use of telephone and face to face modes by using income data of tax authorities. Biemer (1988, 2001) discusses several limitations of this approach, such as unavailability of appropriate external data for all variables of interest, or differences in definition between the survey measurements and the measurements in the external records.

Proxy surveys by which one member of the household (HH) responds for all the other members of the HH are in common use in HH surveys all over the world. The main motivation in this case is to increase the overall sample size, since information is obtained in principle for all the HH members. (Moore, 1988). It also helps in theory to increase the response since if the designated sampled person of the HH cannot be reached or he refuses to respond, another member of the HH is contacted instead. On the other hand, information provided by one member of the HH about another member may be subject to large measurement error (supplying wrong information), and many missing items, ("I don't know"). There seems to be a common perception that proxy-response is less accurate than self-response, (Groves et al. 2004). Kalsbeek et al. (2007) mention a possible cognitive basis for the better quality of self-response over proxy response. There are, however examples where proxy responses turned out to be more accurate, see e.g., O'Muircheartaigh (1991) and also Table 8 in Section 8 of the present paper.

Finally, we note that there exists an ethical problem with the use of proxy surveys, especially in non-mandatory surveys with no obligation to respond. Have the other members of the HH authorized the responding person to provide all the (possibly sensitive) information about them?

In this article we propose to deal with proxy surveys by considering them as a special case of mode effect with the two main modes defined as “direct response” and “indirect response”, where direct response defines that the person provides information about himself and indirect response defines that the response is obtained by another member of the HH. Within each of the two main modes other modes can be defined, like the mode of response, known characteristics of the responding unit, and nonresponse, when no information is obtained from any member of the HH. See Section 8 for an example.

In the following sections we propose and illustrate a new model-based methodology for dealing with mode effects, which does not require a-priori knowledge of a mode providing unbiased estimators. We consider the case of not missing at random (NMAR) nonresponse, by allowing the mode selection probabilities and the probability of nonresponse to depend on the true variable of interest (unobserved under measurement effects) and other explanatory variables. These two parts of our model, the model for the true values and the model for the mode selection account for *selection effects,* with NMAR nonresponse. Nonresponse is considered as another mode. As stated before, ignoring the NMAR nonresponse already induces bias to the sample estimates even in the absence of measurement effects, i.e., when the responses are correct. In order to account for *measurement effects*, we further extend our model by modelling the observed responses as a function of the true target variable, the mode selected and known covariates. Note again that with the existence of measurement effects, the true values of the target variable are unknown. To the best of our knowledge, our approach has not been proposed in the literature. Furthermore, when the covariates are known for all the population values from a census or another register, our approach is applicable also for nonprobability samples.

To fit our three-part model we follow the frequency-based approach with the likelihood maximized by application of an EM algorithm. We discuss converges properties of the algorithm and develop the asymptotic properties of the resulting maximum likelihood estimators. Having estimated all the unknown model parameters, we use the estimated model for predicting the population target quantities. We illustrate our approach by use of simulated data and two real samples, for which the true population values of interest are actually known.

In Section 2, we introduce some notation and define our 3- part model. In Section 3 we describe the estimation of the unknown model parameters and discuss their properties, which are proved in the Appendix at the end of the article. Section 4 considers the estimation of the population parameters of interest, distinguishing between the case where the covariates are known for all the population values of interest and the case where they are known for only the sampled units. Model evaluation is considered in Section 5. In Section 6 we illustrate our approach by simulation experiments, followed by two applications with real data in Sections 7 and 8, with Section 7 focusing on mode-effects and Section 8 on a proxy survey. We conclude with a short summary in Section 9.

2. MODELS FOR SELECTION AND MEASUREMENT EFFECTS

Consider a finite population U of size N and denote by $(Y_i, M_i, \mathrm{X}_i, \mathrm{Z}_i)$ the true outcome variable Y, the response mode M, the auxiliary variables (covariates) X explaining the variability of Y and the covariates Z explaining the variability of M, corresponding to unit i belonging to a sample S of size n, selected from U. In this article we consider the case where Y is binary, taking the values 0 and 1. Suppose first that no measurement effects exist such that every respondent reports his true outcome. Denote by $\overleftrightarrow{M}$ the number of available modes, with the last mode defining the subsample of non-respondents for which only the covariates are known. For convenience, we assume noninformative sampling as defined in Pfeffermann and Sverchkov (1999), such that $\Pr(Y_i \mid \mathrm{X}_i, i \in S) = \Pr(Y_i \mid X_i)$, but the nonresponse is allowed to be NMAR in the sense that $\Pr(\mathrm{R}_i = 1 \mid Y_i, \mathrm{X}_i, i \in S) \neq \Pr(\mathrm{R}_i = 1 \mid \mathrm{X}_i, i \in S)$, where $\mathrm{R}_i = 1$ if sampled unit i responds and $R_i = 0$ otherwise. We further assume that the mode selection depends not only on the covariates Z, but also on the true outcome Y, in the sense that $\Pr(M_i \mid Y_i, \mathrm{Z}_i) \neq \Pr(M_i \mid \mathrm{Z}_i)$. Defining $W_i = \mathrm{X}_i \cup \mathrm{Z}_i$,

$$\Pr(Y_i \mid W_i, M_i) = \frac{\Pr(M_i \mid Y_i, \mathrm{Z}_i)}{\Pr(M_i \mid W_i)} \Pr(Y_i \mid \mathrm{X}_i), \qquad (2.1)$$

where $\Pr(M_i \mid W_i) = \sum_{j=0}^{1} \Pr(M_i \mid Y_i = j, \mathrm{Z}_i) \Pr(Y_i \mid \mathrm{X}_i)$. $\Pr(Y_i \mid \mathrm{X}_i)$ is our tar**g**et population distribution of Y before sampling. It follows from (2.1) that unless M_i is independent of the outcome in the sense that $\Pr(M_i \mid Y_i, \mathrm{Z}_i) = \Pr(M_i \mid \mathrm{Z}_i)$, a mode selection effect is present and with the existence of NMAR nonresponse, the mode effects cannot be ignored in the inference process.

In our empirical study we assume a logistic model for $\Pr(Y_i \mid X_i)$, and a multivariate logistic model for $\Pr(M_i \mid Y_i, \mathrm{Z}_i)$, with 4 possible values for M_i, including nonresponse.

So far, we assumed no measurement effects. Denote by y_i the value measured for responding unit i, which in the case of measurement effects may differ from the true outcome Y_i. We account for possible measurement effects by extending the model (2.1) as,

$$\Pr(y_i \mid W_i, M_i) = \sum\nolimits_{j=0}^{1} \Pr(y_i \mid Y_i = j, W_i, M_i) \frac{\Pr(M_i \mid Y_i, \mathrm{Z}_i)}{\Pr(M_i \mid W_i)} \Pr(Y_i \mid \mathrm{X}_i) . \quad \text{(2.2a)}$$

Denoting, $D_i = I(y_i = Y_i)$ where $I(\cdot)$ is the indicator function, and defining $0^0 = 1$, Equation (2.2a) can be written alternatively as,

$$\Pr(y_i = k \mid W_i, M_i) = \sum\nolimits_{j=0}^{1} \Pr(D_i = j^k (1-j)^{1-k} \mid Y_i = j, W_i, M_i) \Pr(Y_i = j \mid W_i, M_i) . \quad \text{(2.2b)}$$

Application of (2.2b) requires modelling additionally $\Pr(D_i = 1 \mid Y_i = j, W_i, M_i)$ and in our empirical study we again assume a logistic model. Notice that unlike in (2.1), we now only observe the pair (y_i, M_i) and for the non-respondents, only the mode. The target distribution remains $\Pr(Y_i \mid \mathrm{X}_i)$.

3. MODEL ESTIMATION

3.1 The case of no measurement effects

As before, consider first the case of no measurement effects. Adding parameter notation, Equation (2.1) takes the form,

$$\Pr(Y_i \mid W_i, M_i; \alpha_0, \beta_0) = \frac{\Pr(M_i \mid Y_i, \mathrm{Z}_i; \beta_0)}{\Pr(M_i \mid W_i; \alpha_0, \beta_0)} \Pr(Y_i \mid \mathrm{X}_i; \alpha_0) , \quad \text{(3.1)}$$

where $\alpha_0 \in \mathrm{A}$ and $\beta_0 \in \mathrm{B}$ are the true parameter vectors. In what follows we assume that the vectors (Y_i, W_i, M_i) are independent, identically distributed (iid) random variables. Denoting $\delta_0 = (\alpha_0', \beta_0')' \in \mathrm{A} \times \mathrm{B} \equiv \Delta \subset \Re^k$, a compact parameter set, the (full) log likelihood can be written as,

$$L_n(\delta) = n^{-1} \sum\nolimits_{i=1}^{n} \ell_i(\delta) = n^{-1} \sum\nolimits_{i=1}^{n} \Big[I(M_i < \overleftrightarrow{M}) log \Pr(Y_i, M_i \mid W_i; \delta)$$
$$+ I(M_i = \overleftrightarrow{M}) log \Pr(M_i \mid W_i; \delta) \Big]$$

$$= n^{-1} \sum\nolimits_{i=1}^{n} \Big\{ I(M_i < \overleftrightarrow{M}) log[f_i(\delta)^{Y_i} g_i(\delta)^{1-Y_i}] + I(M_i = \overleftrightarrow{M}) log[f_i(\delta) + g_i(\delta)] \Big\} , \quad \text{(3.2)}$$

where $f_i(\delta) = \Pr(Y_i = 1, M_i \mid W_i; \delta)$ and $g_i(\delta) = \Pr(Y_i = 0, M_i \mid W_i; \delta)$.

There exist many optimization algorithms for maximizing the likelihood defined by (3.2). In our empirical study we applied the Newton-Raphson algorithm, which we describe briefly for later use. Denote, $S_n(\delta) = n^{-1}\sum_{i=1}^{n}\nabla_{\delta i}\ell_i(\delta)$, where $\nabla_{\delta i}$ is the gradient with respect to δ and let $D_n(\delta) = n^{-1}\sum_{i=1}^{n}\nabla_{\delta i}^{(2)}\ell_i(\delta)$, represent the corresponding $k \times k$ Hessian matrix. Starting with some initial value δ^0, the Newton-Raphson recursive algorithm is,

$$\delta^j = \delta^{j-1} + D_n^{-1}S_n(\delta^{j-1}), \quad j = 1, 2... \tag{3.3}$$

The iterations continue until $\| \delta^{j+1} - \delta^j \| < \xi$ for some small positive value ξ, where $\| \cdot \|$ is the Euclidean norm. Denote, $H_1(\alpha) = \Pr(Y \mid X; \alpha)$, $H_2(\beta) = \Pr(M \mid Y, Z; \beta)$, $f_M(\delta) = \Pr(Y = 1, M \mid W; \delta)$, $g_M(\delta) = \Pr(Y = 0, M \mid W; \delta)$ and $C_n(\delta) = n^{-1}\sum_{i=1}^{n}\nabla_{\delta i}\ell_i(\delta)\nabla'_{\delta i}\ell_i(\delta)$. Let, $C_{0n} = C_n(\delta_0)$ and $D_{0n} = D_n(\delta_0)$.

In what follows we define regularity conditions for the identifiability and asymptotic properties of the MLE $\hat{\delta}_n$. For this, we assume that as the total sample size n increases, the number of sampled units in each of the modes, except possibly the nonresponse also increases, in the sense that $n_m \geq q_m n; \sum_{m=1}^{\ddot{M}} n_m = n$, for some fixed constants $\{q_m\}$.

A1: The elements of W are bounded almost surely (a.s.) and the functions $H_1(\alpha)$ and $H_2(\beta)$ are identifiable and positive uniformly over A and B a.s. ($H_1(\alpha)$ is positive uniformly over A if $\mathrm{P}[\min_{\alpha \in \mathrm{A}} H_1(\alpha) > 0] = 1$).

A2: The covariates in X (or in Z) contain at least one continuous variable X_v (or Z_v) not contained in Z (X). The derivative of $H_1(\alpha)$ $[H_2(\beta)]$ with respect to this covariate exists and is positive uniformly over A (B) a.s.

A3: Assuming the existence of X_v as above, if $\alpha \neq \alpha^*$, then $\log H_1(\alpha) - \log H_1(\alpha^*) = h_1(X, \alpha, \alpha^*)$ satisfies $\partial h_1(\cdot) / \partial X_v \neq 0$ a.s. Similarly, Assuming the existence of Z_v, if $\beta \neq \beta^*$, then

$\log H_2(\beta) - \log H_2(\beta^*) = h_2(Z,\beta,\beta^*)$ satisfies $\partial h_2(\cdot)/\partial Z_v \neq 0$ a.s. The functions $f_{\ddot{M}}(\delta)$ and $g_{\ddot{M}}(\delta)$ are linearly independent.

A4: δ_0 is an interior vector of Δ and D_{0n} is positive definite for sufficiently large n.

The first three conditions are needed to prove the identifiability and strong consistency of $\hat{\delta}_n$, the maximum likelihood estimator (MLE) of δ. A similar condition to A2 is used in Follmann and Lambert (1991) for the case of a mixture of logistic models with constant mixing probabilities, and in Pfeffermann and Landsman (2011) for modelling non-ignorable assignments in observational studies. The last condition is needed for showing that $\hat{\delta}_n$ is $\sqrt{n}$ consistent and asymptotically normal (CAN). Note that for the logit and probit functions, the conditions A1, A3 and the second part of A4 hold, if the elements of W are linearly independent.

In what follows, $\to_{a.s.}$ defines convergence almost surely and $\to_D$ defines convergence in distribution.

Theorem 1. Under the conditions (A1)-(A3), **(i)** The likelihood (3.2) is identifiable and $\hat{\delta}_n \to_{a.s.} \delta_0$. If, in addition, (A4) also holds, then **(ii)** $C_n^{-1/2} D_n \sqrt{n}(\hat{\delta}_n - \delta_0) \to_D N(0, I_k)$ where I_k is the identity matrix of order k.

3.2 Model with measurement effects

Next consider the model defined by (2.2), which accounts also for measurement effects, in which case the observed measurement y, may differ from the true outcome, Y. Denoting as before $D_i = I(y_i = Y_i)$, we may write,

$$\Pr(y_i, M_i \mid D_i, W_i; \delta) = \begin{cases} \Pr(Y_i, M_i \mid W_i; \delta_0) & \text{if} \quad D_i = 1 \\ \Pr(1 - Y_i, M_i \mid W_i; \delta_0) & \text{if} \quad D_i = 0 \end{cases}, \qquad (3.4)$$

where $\delta_0 \in \mathrm{A} \times \mathrm{B} = \Delta$ is the true parameter vector.

Notice that the models for $\Pr(Y_i \mid \mathrm{X}_i; \alpha_0)$ and $\Pr(M_i \mid Y_i, \mathrm{Z}_i; \beta_0)$ are unchanged but we need to model,

$$p_{ij}(\gamma_0) = \Pr(D_i = 1 \mid Y_i = j, W_i, M_i; \gamma_0), \;\; M_i \neq \overleftrightarrow{M}, \;\; j = 0,1, \qquad (3.5)$$

where $\gamma_0 \in \Gamma$ is the true parameter vector.

Denote $\theta = (\delta', \gamma')'$ and $\theta_0 \in \Theta = \Delta \times \Gamma \subset \Re^s$, a compact set where $s = k + \dim(\gamma_0)$. By (3.4) and (3.5), the log-likelihood is now given by,

$$\tilde{L}_n(\theta) = n^{-1}\sum\nolimits_{i=1}^{n} \tilde{\ell}_i(\theta) = n^{-1}\sum\nolimits_{i=1}^{n} \Big[I(M_i < \overleftrightarrow{M}) y_i \{\log[p_{i1}(\gamma) f_i(\delta) + (1 - p_{i0}(\gamma)) g_i(\delta)]\}$$

$$+(1 - y_i)\{\log[p_{i0}(\gamma) g_i(\delta) + (1 - p_{i1}(\gamma)) f_i(\delta)]\} + I(M_i = \overleftrightarrow{M}) \log[f_i(\delta) + g_i(\delta)] \Big]. \; (3.6)$$

For $M_i \neq \overleftrightarrow{M}$, denote $h_3(\gamma) = (h_3^0(\gamma), h_3^1(\gamma))'$; $h_3^j(\gamma) = \Pr(D = 1 \mid Y = j, W, M; \gamma)$ $j = 0,1$ and define $\tilde{C}_n(\theta)$ and $\tilde{D}_n(\theta)$ in a similar manner as for the model with no measurement effects but with $\tilde{\ell}_i(\theta)$ replacing $\ell_i(\delta)$. Let $\tilde{C}_{0n} = \tilde{C}_n(\theta_0)$, $\tilde{D}_{0n} = \tilde{D}_n(\theta_0)$ and $\hat{\theta}_n$ the MLE maximizing the likelihood.

Suppose the following regularity conditions:

B1: Conditions (A1)-(A3) in Section 3.1 hold, the functions $f_M(\delta)$ and $g_M(\delta)$ are linearly independent also for $M_i \neq \overleftrightarrow{M}$ and $h_3(\gamma)$ is identifiable over Γ.

B2: Condition (A4) holds; γ_0 is interior to Γ and $\tilde{D}_{0n}$ is positive definite for sufficiently large n.

Theorem 2. Under B1, **(i)** The likelihood (3.6) is identifiable and $\hat{\theta}_n \rightarrow_{a.s.} \theta_0$. If, in addition, B2 also holds then **(ii)**, $\tilde{C}_{0n}^{-1/2} \tilde{D}_{0n} \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I_s)$.

3.3 The EM algorithm

To maximize the likelihood (3.6), we apply the iterative EM algorithm (Dempster et al. 1977) which, as shown below, is particularly convenient in the present context. The algorithm has been developed for computing the MLE in cases of incomplete data, which is what happens in our case in the presence of measurement effects, where the true outcomes are unknown.

The key idea underlying the EM algorithm is to add latent variables to the observed data and define a modified likelihood as a function of the observed data and the values of the latent variables. Denote by θ^l the parameter estimates at the l-th iteration. The algorithm cycles between two states. In the first state, it calculates the expected value of the latent variables, denoted by $S_i(\theta^l)$, given the observed data and θ^l. Next, the latent variables in the modified likelihood are replaced by their expected values, thus resulting in a "likelihood" denoted by $M(\theta,\theta^l)$, which depends on $S_i(\theta^l)$ and is much easier to optimize than the original log likelihood (3.6). This step is called the estimation- E-step. In the second state, the likelihood $M(\theta,\theta^l)$ is maximized with respect to the unknown parameters, yielding the estimate, θ^{l+1}. This step is called the maximization- M-step.

In order to implement the algorithm in our case, we define $D_i = I(y_i = Y_i)$ to be the latent variable values. By (3.5)-(3.6), the modified ith log-likelihood is,

$$\begin{aligned}\tilde{\tilde{\ell}}_i(y_i, M_i, D_i \mid W_i;\theta) &= I(M_i < \overleftrightarrow{M})\log[\left(p_{i1}(\gamma)\, f_i(\delta)\right)^{D_i y_i}\left((1-p_{i0}(\gamma))g_i(\delta)\right)^{(1-D_i)y_i} \\ &\times\left(p_{i0}(\gamma)g_i(\delta)\right)^{D_i(1-y_i)}\left((1-p_{i1}(\gamma))\, f_i(\delta)\right)^{(1-D_i)(1-y_i)}] + I(M_i = \overleftrightarrow{M})\log(f_i(\delta)+g_i(\delta))\end{aligned} \quad (3.7)$$

Taking the expectation of D_i given (y_i, M_i) with $\theta = \theta^l$ yields,

$$S_i(\theta^l) = \Pr(D_i \mid y_i, M_i, W_i;\theta^l) = \frac{\Pr(D_i, y_i \mid M_i, W_i;\theta^l)}{\Pr(y_i \mid M_i, W_i;\theta^l)}$$

$$= \frac{[p_{i1}(\gamma^l)\, f_i(\delta^l)]^{y_i}[p_{i0}(\gamma^l)g_i(\delta^l)]^{(1-y_i)}}{\left[\left(p_{i1}(\gamma^l)\, f_i(\delta^l)\right)+\left((1-p_{i0}(\gamma^l))g_i(\delta^l)\right)\right]^{y_i}[p_{i0}(\gamma^l)g_i(\delta^l)+(1-p_{i1}(\gamma^l))\, f_i(\delta^l)]^{(1-y_i)}}. \quad (3.8)$$

This defines the E-step. For defining the M-step, denote the ith log likelihood in (3.2) for the case of no measurement effects as $\ell_i(A_i^{(1)}, A_i^{(2)}, M_i, W_i, \delta^l) = \ell_i(\delta^l)$, where $A_i^{(1)} = Y_i$ and $A_i^{(2)} = 1 - Y_i$. By (3.7)-(3.8) and some simple calculations,

$$\begin{aligned} M(\theta,\theta^l) &= \textstyle\sum_{i=1}^n E\tilde{\tilde{\ell}}(y_i, M_i, D_i \mid W_i, \theta) \\ &= \textstyle\sum_{i=1}^n \{\ell_i[A_i^{(1)}(\theta^l), A_i^{(2)}(\theta^l), M_i, W_i, \delta] + \widehat{\ell}_i(\gamma;\theta^l)\}\end{aligned}, \quad (3.9)$$

where,

$$\widehat{\ell}_i(\gamma;\theta^l) = I(M_i < \overleftrightarrow{M})\ln p_{i1}(\gamma)^{y_i S_i}(1-p_{i0}(\gamma))^{y_i(1-S_i)}\, p_{i0}(\gamma)^{(1-y_i)S_i}(1-p_{i1}(\gamma))^{(1-y_i)(1-S_i)},$$

$$A_i^{(1)}(\theta^l) = y_i S_i + (1-y_i)(1-S_i),\ A_i^{(2)}(\theta^l) = y_i(1-S_i) + (1-y_i)S_i;\ S_i = S_i(\theta^l).$$

The "likelihood" (3.9) is seen to be the sum of two terms, one of only δ and the other of only γ. Thus, the maximization over θ is obtained by maximizing each term separately, simplifying the maximization process substantially.

In our empirical study we set the initial values by drawing many values from a broad uniform prior distribution around $\theta^0 = (\hat{\delta}_n^A, 0)$, where $\hat{\delta}_n^A$ is the MLE of the model without measurement effects, and use the value that maximizes the log-likelihood as the starting initial value.

4. PREDICTION OF FINITE POPULATION MEANS

Replacing the unknown model parameters by their MLE permits estimating the true population mean (proportion), $\bar{Y}_{(P)} = \sum_{j=1}^{N} Y_j$. Denote,

$$\hat{\rho}_i = \Pr(Y_i = 1 \mid \mathrm{x}_i; \hat{\alpha});\ \hat{\tau}_{1,im} = \Pr(Y_i = 1, M_i = m \mid W_i; \hat{\delta}\ ),\ \hat{\tau}_{2,im} = \Pr(M_i = m \mid W_i; \hat{\delta}), \tag{4.1}$$

where $\hat{\delta}' = (\hat{\alpha}', \hat{\beta}')$ denotes the MLE. For the case where the covariates $\{X_i\}$ are known for all the population units, we estimate,

$$\hat{\bar{Y}}_{(Model)} = N^{-1} \sum_{i=1}^{N} \hat{\rho}_i\ . \tag{4.2}$$

Remark 1*.* As long as the model assumed for the population values holds also for the sample, the use of (4.2) does not require probability sampling. Furthermore, if the sample is deemed to be informative in the sense that the inclusion in the sample depends on the true outcome variable, one may extract the population model from the model holding for the sample data, by use of the relationship between the population and sample models, as developed in Pfeffermann and Sverchkov (1999). This requires to model also the probability $\Pr(i \in S \mid Y_i, X_i; \eta)$.

When the covariates are known for only the sampled units and the population size is unknown, we may use a modification of the the Horvitz-Thompson (HT) estimator, i.e.,

$$\hat{\bar{Y}}_{(HT,Model)} = \sum_{i=1}^{n} \pi_i^{-1} \hat{\rho}_i \ / \sum_{i=1}^{n} \pi_i^{-1}\ . \tag{4.3}$$

Assuming that each population unit can be classified by his or her preferred mode, we can also estimate the population mean for each mode, which as discussed in the introduction, is often of great interest on its own. For the case where W_i is known for every $j \in U$, we estimate,

$$\hat{\bar{Y}}_{(Pm)} = \hat{N}_m^{-1}\sum\nolimits_{i=1}^{N}\hat{\tau}_{1,im};\ \hat{N}_m = \sum\nolimits_{i=1}^{N}\hat{\tau}_{2,im}\ . \qquad (4.4)$$

Otherwise,

$$\hat{\bar{Y}}_{Pm,HT} = \hat{N}_{m,HT}^{-1}\sum\nolimits_{i=1}^{n}\pi_i^{-1}\hat{\tau}_{1,im};\ \hat{N}_{m,HT} = \sum\nolimits_{i=1}^{n}\pi_i^{-1}\hat{\tau}_{2,im}\ . \qquad (4.5)$$

It is easy to show that under the conditions of Theorem 2, the predictors defined by (4.1)-(4.4) are $\sqrt{n}$-consistent for the corresponding true population means, under an appropriate asymptotic framework for finite population sampling. See Isaki, and Fuller (1982) for such a framework.

Remark 2. The predictors (4.1)-(4.5) are computed the same way under both the models with, and without measurement effects.

5. MODEL EVALUATION

5.1 The Hosmer-Lemeshow test

The predictors developed in the previous sections are model-dependent and as such, the model used for their construction needs to be tested. Many goodness-of-fit test procedures under the frequentist approach for continuous outcomes have been proposed in the literature. See, e.g., Pfeffermann and Landsman (2011) and Pfeffermann and Sikov (2011) for review and references. In our empirical study we consider the case of binary outcomes generated from logistic models and we apply the Hosmer Lemeshow (HL, 1980, 2000) goodness-of-fit test, which is in common use for testing models for binary data.

The test statistic compares within pre-specified groups the number of observed successes ($y_i = 1$), with the number of expected successes, as predicted under the estimated model. For this, the data are first ordered according to the predicted probability of success under the model evaluated. Next, the units are grouped based on the ordering, into a certain number of groups of approximately equal size, ($G = 10$ groups is common), and within each group the estimated expected number of successes, (the sum of the predicted probabilities of success), is compared to the observed number of successes. The test statistic is,

$$HL = \sum\nolimits_{g=1}^{G} \frac{(O_g - n_g \bar{\hat{p}}_g)^2}{n_g \bar{\hat{p}}_g (1 - \bar{\hat{p}}_g)} \sim \chi^2_{(G-2)}, \tag{5.1}$$

where O_g is the number of observed successes in group g, n_g is the number of units in the group and $\bar{\hat{p}}_g = n_g^{-1} \sum\nolimits_{i=1}^{n_g} \hat{p}_{gi}$ is the mean of the estimated probabilities of success. Hosmer and Lemeshow (1980) found through empirical studies under a much simpler setup than in our case that the test statistic (5.1) follows approximately the χ^2 distribution with $(G-2)$ degrees of freedom under the null hypothesis that the model fits the data.

5.2 Normalized likelihood ratio (N-LR) test for model selection

A standard test of the null hypothesis that two models, one nested within the other, fit the data "equally well", is the likelihood ratio test. We apply the test (with certain correction, see below), for testing the null hypothesis that accounting for measurement effects in the extended model (2.2) does not improve the goodness-of-fit compared to the model (2.1) with only selection effects, or more formally, for testing that there are no measurement effects.

Distinguishing the models without and with measurement effects by the superscripts *A* and *B* respectively, the standard test statistic is,

$$LR_n = -2[l_n^A(\hat{\delta}) - l_n^B(\hat{\theta})], \tag{5.2}$$

where $l_n^A(\hat{\delta}) = \sum\nolimits_{i=1}^{n} \ell_i^A(\hat{\delta})$ and $l_n^B(\hat{\theta}) = \sum\nolimits_{i=1}^{n} \ell_i^B(\hat{\delta})$ are the corresponding log-likelihoods computed with the MLEs, as obtained under the two models. (Equations 3.2 and 3.6). However, unlike in standard problems where the nested model is obtained from the extended model by nullifying certain parameters, this is not the case in our application, where we restrict to logistic models, since $[1+\exp(-t)]^{-1} \neq 0$ for all t in a compact set and hence the model A without measurement effects is not nested in the model B with them. To deal with this problem, Vuong (1989) proposed a normalized likelihood ratio test (N-LR) defined as,

$$LR_{nor} = \frac{LR_n}{\hat{\omega}_n}, \tag{5.3}$$

where $\hat{\omega}_n^2 = 4n\left[n^{-1}\sum_{i=1}^{n}[l_i^A(\hat{\delta}) - l_i^B(\hat{\theta})]^2 - (n^{-1}LR_n)^2 \right]$ is an estimator of the variance of LR_n. The author shows that under the assumption that the true parameter vector is an interior point and some other regularity conditions, the asymptotic distribution of LR_{nor} is normal. However, when model A is correct, one would expect that some of the parameters γ which define the probability of the measurement effects lie on the boundary of the assumed parameter set. In this case, Vuong's condition that the true parameter vector is an interior point is violated. A similar problem is demonstrated in Wilson (2015). In the empirical study we approximated the distribution of the statistic LR_{nor} by parametric bootstrap.

6. EMPIRICAL RESULTS BASED ON SIMULATIONS

6.1 Design of simulation study

In order to assess the performance of our proposed approach, we designed a simulation study which consists of the following sages:

1. Generate a population of size $N = 10^5$ with values of three covariates, X_{1i}, X_{2i}, X_{3i}, generated independently from a $Beta(2,5)$ distribution.
2. Generate a binary outcome Y_i from the logistic distribution;

$$\Pr(Y_i = 1 | X_{1i};\alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_1 X_{1i}),\ i = 1,...,N\ .$$

(6.1)

3. Classify the population units into four modes with probabilities,

$$\Pr(M_i = m | Y_i, X_{2i}; \beta_m) = \frac{\exp(\beta_{0m} + \beta_{1m}X_{2i} + \beta_{2m}Y_i)}{1 + \sum_{m=1}^{3}\exp(\beta_{0m} + \beta_{1m}X_{2i} + \beta_{2m}Y_i)},\ m = 1,2,3$$
$$\Pr(M_i = 4 | Y_i, X_{2i}; \beta_m) = 1 - \sum_{m=1}^{3}\Pr(M_i = m | Y_i, X_i; \beta_m), \tag{6.2}$$

where $M_i = 4 = \overleftrightarrow{M}$ defines the "mode" of nonresponse. Notice that the model assumes NMAR nonresponse as the probability not to respond depends directly on the outcome. The parameter values in (6.1)-(6.2) were set such that the relative population sizes in the 4 modes are approximately (10%, 25%, 40%, 25%).

The population in Stages 1-3 has been generated only once, such that the assessment of the performance of the proposed methodology is "design-based", over all possible sample selections from a fixed population. Alternatively, one could generate many populations and draw samples from each population, but with a population of size $N = 10^5$, this would not make much difference.

4. Draw $K = 1000$ samples of size $n = 5000$ by simple random sampling without replacement from the population obtained in 1-3.

5. Generate measured values for the sampled units from the model,

$$\Pr(D_i = 1 \mid Y_i, M_i = m, X_{3i}; \gamma_m) = \text{logit}^{-1}(\gamma_{0m} + \gamma_{1m} X_{3i} + \gamma_{2m} Y_i)\ ,\ m = 1,2,3 \qquad (6.3)$$

6. Estimate the model coefficients for each sample by maximizing the likelihood function for the respective model (with or without measurement effects). We used the Newton Raphson algorithm (3.3) for estimating the coefficients of the model without measurement effects (hereafter model A), and the EM algorithm described in section (3.3) for the model with measurement effects (hereafter model B). We used parametric bootstrap for estimating the standard errors (S.E) of the mean estimators with 100 bootstrap samples for each parent sample. Estimating the S.E. of the model coefficients by use of the inverse information matrix turned out to be unstable, with occasional very extreme and even negative variance estimators. The computer code for running the simulation study (and for the applications with real data in Section 7) has been written in MATLAB version 9.5.

6.2 Simulation results

Table 1 contains the mean of the estimated coefficients (Mean Est.), along with their empirical S.E., for the models without (model A) and with (model B) measurement effects. The empirical S.E. are the standard deviations of the estimates over the 1,000 samples divided by $\sqrt{1,000}$. We also computed the means of the estimated bootstrap S.E. but they are not shown since they are very close to the empirical S.E. for all the coefficients, under both the models

A and B. We note in this respect that the measurement effects under Model B are not negligible. We computed for the first of the 1,000 samples the probabilities of measurement effects; $\mathrm{a}_i = \Pr(D_i = 0 \,|\, Y_i, X_{2i} X_{3i})$ and found that Min(a_i)=0.01, Q1(a_i)=0.12, Mean (a_i)= 0.36), Q3(a_i)=0.59, Max(a_i)=0.99. (Q1 and Q3 are the 1st and 3rd quarters.)

Table 1. True coefficients, means of estimated coefficients and empirical standard deviations (S.E.) of estimates. 1000 samples.

Modes	True Coefficients	Model A		Model B	
		Mean Est.	Emp. S.E.	Mean Est.	Emp. S.E.
	$\alpha_0 = 1$	0.99	0.004	1.01	0.01
	$\alpha_1 = -1$	-0.99	0.003	-1.04	0.09
$m = 1$	$\beta_{01} = -0.5$	-0.53	0.011	-0.53	0.012
	$\beta_{11} = 3$	3.00	0.014	2.90	0.026
	$\beta_{21} = -1$	-0.99	0.006	-0.97	0.008
$m = 2$	$\beta_{02} = 1$	1.08	0.009	1.04	0.009
	$\beta_{12} = 3$	2.98	0.012	2.89	0.026
	$\beta_{22} = -2$	-2.00	0.007	-2.01	0.008
$m = 3$	$\beta_{03} = 1$	0.98	0.008	1.01	0.008
	$\beta_{13} = 3$	3.02	0.013	2.95	0.027
	$\beta_{33} = -3$	-3.01	0.008	-3.01	0.009
$m = 1$	$\gamma_{01} = 0$	--	--	0.02	0.026
	$\gamma_{11} = -1$	---	---	-1.01	0.022
	$\gamma_{21} = 0.5$	---	---	0.55	0.030
$m = 2$	$\gamma_{02} = 0.5$	---	---	0.45	0.015
	$\gamma_{12} = -1$	---	---	-1.01	0.011
	$\gamma_{22} = 1$	---	---	1.02	0.019
	$\gamma_{03} = 0.3$	---	---	0.29	0.009

$m=3$	$\gamma_{13}=-1$	---	---	-0.97	0.009
	$\gamma_{23}=1.5$	---	---	1.48	0.017

As seen in Table 1, the mean estimates under both models are very close to the corresponding true coefficients with very small standard errors, although it should be noted that in most cases the differences are significant when tested by use of the standard *t*-statistic. As expected, the empirical S.E. under Model *B* are somewhat higher than the corresponding S.E. under Model *A*, as results from the existence of measurement effects under Model B. As mentioned above, the means of the bootstrap S.E. estimators (not shown) are very close to the corresponding empirical S.E.

To save in space, in what follows we restrict to only to the results obtained under Model B with the measurement effects. Similar (and somewhat better) results have been obtained under Model A for which the observed values are the same as the true values with no errors.

Table 2 shows the results obtained when predicting the population means and sizes for the various modes, using the predictors defined by (4.2)-(4.5). The predictor $\bar{y}_S$ is the HT estimator when ignoring the mode effects, using all the measurements including for the non-respondents. (Reduces to the simple sample mean based on all the sample units under simple random sampling).

Table 2. Mean predictions of population and sub-population means and sizes, and empirical S.E. (in parenthesis). 1000 samples.

Predictor	$m=1$	$m=2$	$m=3$	$m=4$
N_m	8262	24017	44933	22788
$\hat{N}_m$	8277 (32)	24053 (55)	44989 (68)	22760 (41)
$\hat{N}_{m,HT}$	8240 (34)	24040 (53)	44960 (68)	22700 (53)
$\bar{Y}_{Pm}$	0.45	0.60	0.66	0.87
$\hat{\bar{Y}}_{Pm}$	0.44 (0.0035)	0.60 (0.0020)	0.65 (0.0015)	0.87(0.0013)
$\hat{\bar{Y}}_{Pm,HT}$	0.44 (0.0037)	0.61 (0.0022)	0.65 (0.0015)	0.86 (0.0015)

$\bar{Y}_{(P)} =$ 0.67	$\hat{\bar{Y}}_{(Model)} =$ 0.67 (0.001)	$\hat{\bar{Y}}_{(HT,Model)} =$0.67 (0.001)	$\bar{y}_S = 0.62$ (0.002)

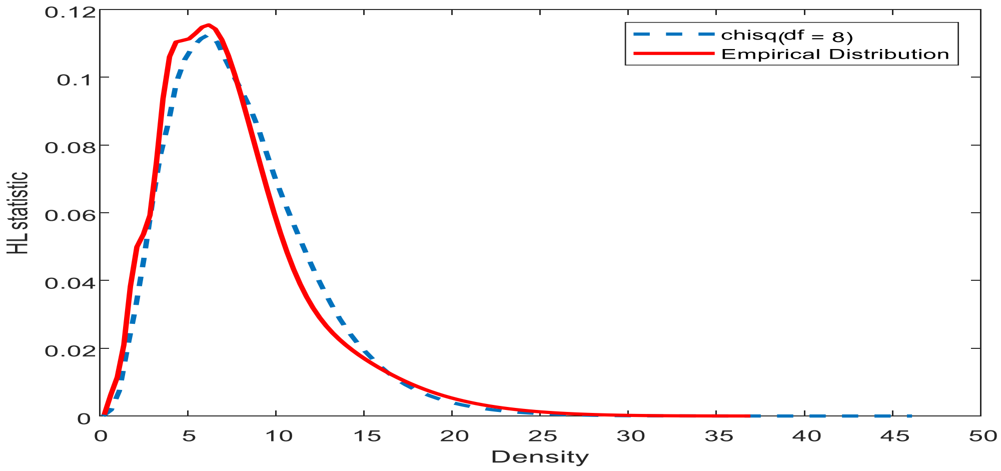
Table 2 shows very good performance of the predictors, despite the existence of nonignorable measurement effects. The HT estimators have slightly larger S.E. than the model-based predictors, which of course is expected since the latter predictors use the information about the population covariates, not used by the HT estimator. Notice how well the size and true mean of the non-respondents ($m = 4$) are predicted, even with only the HT estimator. Finally, the use of the estimator $\bar{y}_S$ (simple sample mean in our case), which ignores the mode effects is seen to be biased, illustrating that mode effects cannot be ignored when estimating concurrent population means. (When ignoring also the nonresponse and computing the mean of only the observed measurements, the corresponding estimator is $\bar{y}_S = 0.52$.)

6.3 Model evaluation

6.3.1 *Distribution of Hosmer-Lemeshow statistic under model B*

To illustrate the distribution of the Hosmer-Lemeshow (HL) we drew 1,000 new samples of size 5,000 without replacement from the same population and subjected the true outcomes of the responding units to measurement effects via the model (6.3). Figure 1 shows the smoothed empirical density of the test statistic for $G = 10$ nearly equal-size groups over the 1,000 samples. Recall that the HL statistic is supposed to have a ${\chi^2}_{(8)}$ distribution under correct model specification (Model B in our case). As can be seen, the two densities are indeed very close, supporting the conjecture that the true distribution is indeed ${\chi^2}_{(8)}$.

Figure 1. Empirical distribution of HL statistic under model B. Nonparametric density estimation. Comparison to $\chi^2(8)$ density.



In order to study the power of HL statistic, we assume under the null hypothesis that the model for the mode selection probabilities is as defined by (6.2), where in fact we generated the modes using the model,

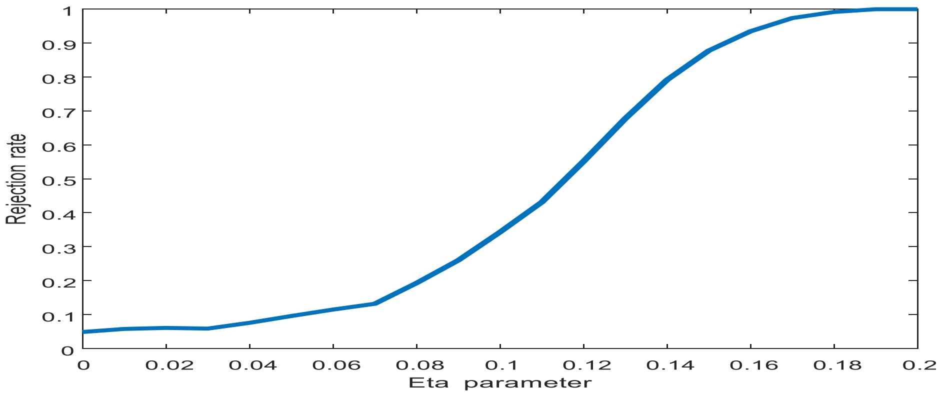
$$\Pr(M_i = m \mid Y_i, X_{2i}; \beta_m, \eta) = \frac{\exp(\beta_{0m} + \beta_{1m} Y_i + \beta_{2m} X_{2i} + \eta X_{2i}^2)}{1 + \sum_{j=1}^{M} \exp(\beta_{0m} + \beta_{1m} Y_i + \beta_{2m} X_{2i} + \eta X_{2i}^2)}, \quad m = 1,2,3 \tag{6.4}$$

$$\Pr(M_i = 4 \mid Y_i, X_{2i}; \beta_m, \eta) = 1 - \sum_{m=1}^{3} \Pr(M_i = m \mid Y_i, X_{2i}; \beta_m, \eta).$$

(Compare with 6.2).

Figure 2 shows the empirical rejection rates when using the HL test for values η in the range $\eta \in [0, 0.2]$, based on 1000 samples of size $n = 5{,}000$ for each value η, generated by use of (6.4). For $\eta = 0$, (correct model specification), the rejection rate is close to the nominal size of 0.05, as it should be. As η increases (the assumed model is further away from the correct model), the rejection rates increase monotonically, reaching powers ≥ 0.8, already for $\eta \geq 0.15$.

Figure 2. Rejection rates with significance level of 0.05, when true mode selection probabilities are defined by (6.4). 1000 samples, $\eta \in [0, 0.2]$.



6.3.2 *Distribution of the N-LR statistic under the correct model*

The purpose of this section is to illustrate that we can approximate the distribution of LR_{nor}, the N-LR test statistic as defined in Equation 5.3 by parametric bootstrap. For this, we generated 1,000 samples of size 5,000 under Model A (Equations 6.1-6.2) with the same true parameters as before, and other 1,000 samples of size 5,000 under model B. The first set of samples allow us to assess the approximation of the distribution under the null hypothesis of no measurement effects, while the second set allows to assess the distribution of the test when there exist measurement effects. For this, we followed the following steps:

1) Estimate the models A and B for each of the 2,000 samples and compute the LR_{nor} test statistic. At the end of this step we have **1,000** observations of the test statistic with data obeying the null hypothesis of no measurement effects, and 1,000 observations of the test statistic with data containing measurement effects as under the alternative hypothesis, allowing us to estimate the true distributions under the two hypotheses by the corresponding empirical distributions.

2) Generate for each of the first 10 samples generated under Model A, 1000 parametric bootstrap samples of size 5,000 under the model A, using the corresponding estimated coefficients from the parent sample. Re-estimate the two models for each bootstrap sample and compute the LR_{nor} statistic. This

step generates 10,000 realizations form the bootstrap distribution of the test statistic when the null hypothesis of no measurement effects holds.

3) Repeat Step 2 but this time by generating the bootstrap samples from Model B, as estimated for each of the first 10 parent samples generated under this model. This step generates 10,000 realizations form the bootstrap distribution of the test statistic under the alternative hypothesis that measurement effects exist.

Figures 3 and 4 compare the empirical distributions under the two hypotheses with the corresponding bootstrap approximations, showing sufficiently close fit in both cases. Notice that when the actual observed data conform with Model A, it is the first bootstrap distribution which will practically be used for testing the null hypothesis of no measurement effects. When the observed data conform with Model B, it is the second bootstrap distribution which will practically be used.

Figure 3. Empirical and Bootstrap distributions of the normalized LR statistic under Model A.

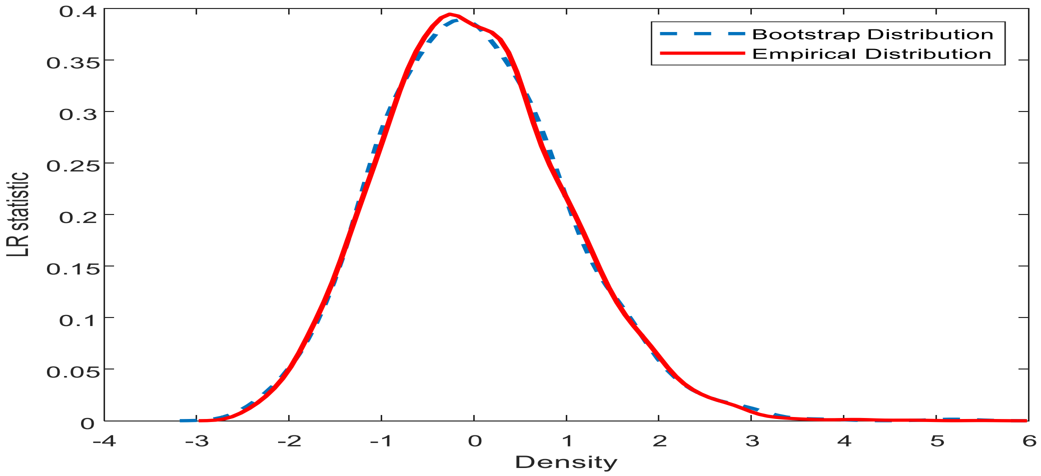
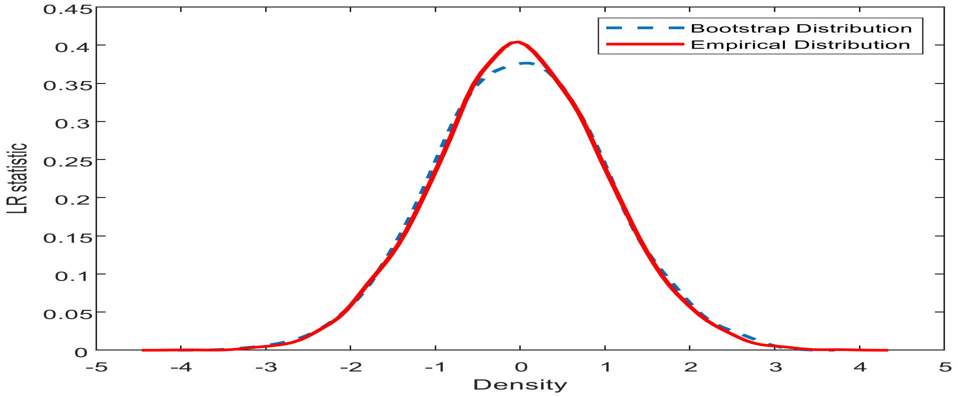


Figure 4. Empirical and Bootstrap distributions of the normalized LR statistic under Model B.



7. EMPIRICAL RESULTS FOR A REAL MIXED-MODE SURVEY

7.1 Example of measurement effects

As explained in the introduction, the term measurement effect refers to the case where a sampled unit responds differently, depending on the mode of response. In the Agriculture Census of Israel carried out in 2018, 210 farmers (out of about 17,000) happened to respond both by Telephone and via the internet, after receiving mistakenly a reminder to respond via the Internet, even though they already responded by telephone. Table 3 summarizes the results obtained for two of the questions asked in the census: number (#) of workers in the farm, and total cultivated area. Out of the 210 farmers, 131 farmers responded the same way on Question 1 and 139 farmers responded the same way on Question 2. The notation T>I (T<I) defines the farmers with the higher (lower) responses on telephone than on the internet.

The figures in the table indicate big differences in the answers of the farmers that provided different answers by the two modes (about one third of the farmers responding by the two modes). In this example the measurement effects cancel out when computing the means, but there is no guarantee that this always happens, and proper models need to be used to account successfully for the measurement effects. See next section.

Table 3. Mean responses obtained by farmers responding both by telephone and internet: overall and separately for farmers with T>I and T<I.

Questions	Mean Internet	Mean Telephone	Mean and # for T>I	Mean and # for T<I
# of workers	5.9	5.8	T=15.5, I= 7.0 39 farmers	T= 7.5, I=17.0 40 farmers
Cultivated area	108.5	105.9	T= 318.4, I= 192.0 38 farmers	T= 88.3, I= 144.5 33 farmers

7.2 Accounting for mode effects in a real mixed-mode survey

In this section we illustrate the performance of our proposed approach by using data collected as part of the annual crime victimization survey in 2017, administered by the Israel Central Bureau of Statistics (ICBS). The survey collects information on victimization with respect to a variety of crimes, as well as socio-demographic information. Similar surveys are carried out by national statistical offices throughout the world. The sample is drawn by probability sampling, and the sampled units can respond using either the internet, or by telephone, implying that we have 3 modes, with the third mode defined by non-respondents. The total sample size is $n = 7035$, with 11% responding via the internet (I), 60% by telephone (T) and 29% not responding (NR).

Although not a primary variable of interest in this survey, we chose as the target outcome variable the binary variable Y_i, taking the value "1" if unit i has an academic degree (Bachelor or higher). The aim is to predict the true population proportion of persons with academic degree. The reason for this choice is that the ICBS has an extensive register of education, with population coverage of over than 95%, so that we can assess the performance of our method by comparing the predictors to the "truth". The register is complete for only 2016, but we don't expect significant differences between the two years. We confined our analysis to persons aged 20+ and based on the register, the true proportion is $\bar{Y}_{(P)} = 0.24$. On the other hand, 41.4% of the internet respondents and 23.5% of the telephone respondents in the sample have an academic degree, indicating the existence of a mode selection effect, and possibly also measurement effects.

In what follows we present and discuss the results of our study. We included in the covariates (X) the variables gender (gen; male=1), age (in years), and country of birth (cob; Israel=1, other countries=0). In accordance with the general methodology outlined in Sections 2-4, and the models and notation in Section 6, we fitted the following models:

$$\Pr(Y_i = 1 \mid X_i; \alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_1 age_i + \alpha_2 cob_i + \alpha_3 gen_i) \tag{7.1}$$

$$\Pr(M = m \mid X_i, Y_i; \beta) = \frac{\exp(\beta_{0m} + \beta_{1m} Y_i + \beta_{2m} cob_i + \beta_{3m} gen_i)}{1 + \sum_{m=1}^{2} \exp(\beta_{0m} + \beta_{1m} Y_i + \beta_{2m} Cob_i + \beta_{3m} gen_i)},\ m = T, I \tag{7.2}$$

$$\Pr(M_i = NR \mid Y_i, X_i; \beta_1, \beta_2) = 1 - \sum_{m=I,T} \Pr(M_i = m \mid Y_i, X_i; \beta_m),$$

$$\Pr(D_i = 1 \mid Y_i, M_i = m, X_i; \gamma_m) = \text{logit}^{-1}(\gamma_{0m} + \gamma_{1m} Y_i + \gamma_{2m}\, age_i + \gamma_{3m} gen_i),\ m = T, I \tag{7.3}$$

Table 4 shows the estimated coefficients and the parametric bootstrap S.E. of the models (7.1) and (7.2), as obtained when fitting the models with- and without accounting for possible measurement effects (M.E.). Table 5 shows the estimated coefficients and S.E. of the model (7.3), which accounts for measurement effects. We also computed the means of the estimated coefficients over all the BS samples (not shown), and found that they are very close to the estimates obtained from the original sample, thus verifying the asymptotic unbiasedness of our ML estimators.

As can be seen, most of the coefficients are highly significant in both tables under the standard t-test. What we find a bit surprising is that the coefficient of Y (having academic degree) in the models for $M_i = T \mid X_i, Y_i; \beta$ and $M_i = I \mid X_i, Y_i; \beta$ is highly negative in all 4 models in Table 4, suggesting that with the other covariates held fixed, having an academic degree actually encourages nonresponse (the third mode). Also, for the model with measurement effects (Table 5), the coefficient of Y is positive and highly significant (but of much smaller magnitude), suggesting that having an academic degree increases the probability of misreporting. Including more covariates in the models could possibly resolve these, somehow unexpected outcomes.

Table 4. Estimation of α and β coefficients (Eqs. 7.1, 7.2), when fitting the model without- and with accounting for possible measurement effects (M.E.). S.E. are based on 1000 parametric bootstrap samples.

Model for	covariates	Model fitted without M.E.		Model fitted with M.E.	
		Est.	S.E	Est.	S.E.
$Y_i \mid X_i;\alpha$ Eq. (7.1)	*const.*	-1.0302	0.0634	-1.2264	0.077
	age	0.0182	0.0126	-0.054	0.0259
	Cob	-0.0938	0.0665	-0.01904	0.0011
	gen.	0.3150	0.0580	0.2669	0.0015
$M_i = T \mid X_i, Y_i;\beta$ Eq. (7.2)	*const.*	2.3671	0.0498	2.3945	0.3701
	Y	-2.1040	0.0428	-1.9865	0.1113
	Cob	0.0669	0.0573	0.0512	0.0456
	gen.	0.0639	0.0566	0.0914	0.0183
$M_i = I \mid X_i, Y_i;\beta$ Eq. (7.2)	*const.*	0.9694	0.0599	1.1701	0.1971
	Y	-2.7375	0.0555	-2.8263	0.2244
	Cob	0.3360	0.0857	0.3476	0.0456
	gen.	-0.1798	0.0898	-0.1307	0.0547

Table 5. Estimation of γ coefficients (Eq. 7.3), when fitting the model which accounts for possible measurement effects. S.E. based on 1000 parametric bootstrap samples.

Model	covariates	Est.	S.E
$D_i \mid Y_i, M_i = T, X_i;\gamma_m$ Eq. (7.3)	*const.*	0.2864	0.0016
	Y	0.2934	0.0017
	age	0.0686	0.0273
	gen.	0.2331	0.0013
$D_i \mid Y_i, M_i = I, X_i;\gamma_m$ Eq. (7.3)	*const.*	0.2827	0.0016
	Y	0.1636	0.0009
	age	0.1017	0.0539

	gen.	0.1845	0.0010

Next, we study the performance of the model in predicting the true population proportion, which as noted in Section 7.1, is basically known for this application. We computed the following predictors:

A- $\hat{\bar{Y}}_{(Model)} = N^{-1}\sum_{i=1}^{N}\hat{\rho}_i$; $\hat{\rho}_i = \Pr(Y_i = 1 | \mathrm{x}_i; \hat{\alpha})$ (uses the covariate information for all the population units). (Equation (4.2)

B- $\hat{\bar{Y}}_{(HT,Model)} = \sum_{i=1}^{n}\pi_i^{-1}\hat{\rho}_i \,/ \sum_{i=1}^{n}\pi_i^{-1}$; $\pi_i = \Pr(i \in S)$ (Equation 4.3)

C- $\hat{\bar{Y}}_{HT,True} = \hat{N}^{-1}\sum_{i=1}^{n}\pi_i^{-1}Y_i^{True}$; $\hat{N} = \sum_{i=1}^{n}\pi_i^{-1}$ (uses the true values Y_i known from the education register).

D- $\hat{\bar{Y}}_{HT,Adj} = \hat{N}_{Adj}^{-1}\sum_{i=1}^{n}\tilde{\pi}_i^{-1}y_i$; $\hat{N}_{Adj} = \sum_{i=1}^{n}\tilde{\pi}_i^{-1}$, the modified HT estimator but with the standard base weights $\{a_i = \pi_i^{-1}\}$ replaced by adjusted weights to account for the nonresponse.

E- $\hat{\bar{Y}}_{HT,imp} = \hat{N}^{-1}\sum_{i=1}^{n}\pi_i^{-1}\tilde{Y}_i$; $\tilde{Y}_i = y_i$ if unit i responds, $\tilde{Y}_i = Y_{i,imp}$ if unit i does not respond. The imputation was carried out using the monotone imputation method of Rubin (1987, p. 172), based on the observed sample values y_i .

Table 6. Predictors of proportion of persons with academic degree. Crime victimization survey, ICBS, 2017.

Measurement effects included(?)	$\bar{Y}_P$ (true)	$\hat{\bar{Y}}_{(Model)}$	$\hat{\bar{Y}}_{HT,Model}$	$\hat{\bar{Y}}_{HT,True}$	$\hat{\bar{Y}}_{HT,Adj}$	$\hat{\bar{Y}}_{HT,imp}$
NO (model A)	0.24	0.26	0.28	0.25	0.36	0.33
YES (Model B)		0.25	0.23			

Computing the Hosmer-Lemeshow (HL) test discussed in Section 5.1 under the two models, and the normalized likelihood ratio (N-LR) test discussed in Section 5.2, yields (p-values in parentheses): $HL(A) = 11.6\ (p\text{-}value = 0.17)$, $HL(B) = 9.44\ (p\text{-}value = 0.31)$, $LR_{nor} = 0.21\ (p\text{-}value = 0.34)$.

The results of this study show very clearly that our proposed model-based predictors are much superior to the design-based estimators, which ignore the mode effects ($\hat{\bar{Y}}_{HT,Adj}, \hat{\bar{Y}}_{HT,imp}$), despite the use of only three covariates for which the population values are known. (The estimator $\hat{\bar{Y}}_{HT,True} = 0.25$, which uses the correct values of the outcome variable indicates that the design-based estimator in the case of no measurement effects and nonresponse performs well.) Model B, which accounts for possible measurement effects seems to perform somewhat better than Model A, which assumes no measurement effects (note the relative high value of $\hat{\bar{Y}}_{HT,Model} = 0.28$ under Model A), although this is only partly reflected by the values of the two test statistics, suggesting that for the variable of having an academic degree in this survey, there are only small measurement effects, not detected by the two tests considered.

8. DEALING WITH PROXY SURVEYS AS MODE EFFECTS

As mentioned in the introduction, we propose dealing with the problem of proxy surveys via the methodology developed in the present article for dealing with mode effects. We illustrate our proposal using data collected as part of the Labor Force Survey (LFS), administered by ICBS. The LFS in Israel is a monthly survey with a 4- in, 8- out, 4- in, rotation pattern. For the present illustration we use the data observed in all the months of 2018 for the first interview, which is carried out by a personal interview. To further reduce the overall sample size, we restrict to the Jewish population aged 20-40, yielding a sample of $n = 19{,}820$ persons.

We again use the binary variable Y_i - "having an academic degree", as our target variable, thus allowing us to compare predictors of the population proportion of people with academic degree to the true proportion. Table 7 shows a few design-based estimators computed from the data, after modifying the base sampling weights to account for nonresponse. The estimators shown are:

$\hat{\bar{Y}}_{NoM\mathrm{odes}}$ - Standard HT estimator when ignoring the mode effects, but with the sampling weights adjusted for non-response,

$\hat{\bar{Y}}_{Direct}$ - Design-based estimator using only the direct responses (with adjusted weights),

$\hat{\bar{Y}}_{\text{Proxy}}$ - Design-based estimator using only the proxy responses (with adjusted weights).

Table 7. Proportion of persons with academic degree. Preliminary design-based estimators. LFS, Jewish population age 20-40, ICBS, 2017.

$\bar{Y}_P$ (true)	$\hat{\bar{Y}}_{No\text{Modes}}$	$\hat{\bar{Y}}_{Direct}$	$\hat{\bar{Y}}_{\text{Proxy}}$
0.248	0.310	0.431	0.268

As expected, the design-based estimator $\hat{\bar{Y}}_{No\text{Modes}}$ that ignores the mode effects performs poorly. The estimator based on only the direct responses performs even worse but quite surprising, the estimator $\hat{\bar{Y}}_{\text{Proxy}}$, which uses only the proxy responses performs relatively well. It seems therefore that when asked about the possession of an academic degree, the proxy responses are generally more accurate than the responses of interviewees responding about themselves. The importance of this outcome is in illustrating that it is not necessarily true that interviewees responding about themselves provide correct answers, or in a more general mode effects set up, that one can decide on a mode with correct answers. Recall from the Introduction that several methods proposed in the literature to deal with mode effects assume the existence (and knowledge) of a mode which provides unbiased estimators for the true population mean.

We now illustrate the use of our mode effects methodology to handle proxy survey problems. We consider 5 different "modes" as follows: proxy response-male (MP), proxy response- female (FP), direct response- male (MD), direct response- female (FD), nonresponse (NR). By MP we mean that the unit for which a proxy response is provided is a male and similarly for the other modes. Out of the total sample size $n = 19{,}820$, 16.8% responses have been obtianed by FD, 15% by MD, 30.1% by FP, 31.8 by MP and 6.3% did not respond. Denoting by S_m the sample of units responding by mode m, we estimated the population proportion for each of the modes using the ratio HT estimator with weights adjusted for nonresponse,

$\hat{\hat{Y}}^{(m)}_{HT,Adj} = \hat{N}^{-1}_{Adj,m}\sum_{i\in S_m}\tilde{\pi}_i^{-1}y_i;\ \hat{N}_{Adj,m} = \sum_{i\in S_m}\tilde{\pi}_i^{-1}$ and found: $\hat{\hat{Y}}^{(MP)}_{HT,Adj} = 0.19$, $\hat{\hat{Y}}^{(FP)}_{HT,Adj} = 0.34$, $\hat{\hat{Y}}^{(\mathrm{MD})}_{HT,Adj} = 0.37$, $\hat{\hat{Y}}^{(\mathrm{FD})}_{HT,Adj} = 0.48$, suggesting the existence of mode effects. (Similar differences exist when using instead the true values of Y as known from the register.)

We use as covariates age and years of study. (Years of study is known from the register, the gender of the interviewee is accounted for in the definition of the modes.) To save in space, we do not present the coefficients of the models (7.1)-(7.3) obtained in this case.

Table 8 shows again the true value and the different predictors obtained in this case. The notation is the same as before.

Table 8. Proportion of persons with academic degree. Model- and design-based estimators. LFS, Jewish population age 20-40, ICBS 2017.

Measurement effects included(?)	$\bar{Y}_P$ (true)	$\hat{\hat{Y}}_{(Model)}$	$\hat{\hat{Y}}_{HT,Model}$	$\hat{\hat{Y}}_{HT,True}$	$\hat{\hat{Y}}_{HT,imp}$	$\hat{\hat{Y}}_{HT,Adj}$
NO (model A)	0.248	0.304	0.306	0.238	0.305	0.305
YES (Model B)		0.252	0.271			

The results in Table 8 show very good performance of the model-based predictors when fitting Model B, which accounts for possible measurement effects, with the estimator $\hat{\hat{Y}}_{(Model)}$ that uses all the population covariates yielding an almost perfect predictor. On the other hand, the estimators obtained under Model A, which assumes no measurement effects are clearly biased, indicating the existence of measurement effects in this application. This result is reinforced by the HL test statistics, rejecting the null hypothesis of Model A, $HL(A) = 26.359\ (p - value = 0.001)$ but not rejecting Model B, $HL(B) = 7.24\ (p - value = 0.51)$. Also, the N-LR test rejects Model A in favor of Model B, $LR_{nor} = 10.31\,(p - value = 0.00)$.

The estimator $\hat{\hat{Y}}_{HT,True}$, which uses the true Y-values from the education register performs very well, validating the sampling design and corresponding estimator, but the estimator $\hat{\bar{Y}}_{HT,imp}$ which imputes the missing data for the nonrespondents based on the observed data and thus ignores the measurement effects, and the estimator $\hat{\bar{Y}}_{HT,Adj}$, which attempts to correct for the nonresponse by modification of the sampling weights (but does not account for the measurement effects) perform poorly, over-estimating the true proportion by about 23%, the same as the model-based estimators under Model A. We conclude that in this application, there are large measurement effects, captured well under Model B, but not under Model A and the modified design-based estimators considered.

9. SUMMARY

In this article we propose a new comprehensive model-based approach to deal with mode-effects, which is applied also to deal with proxy surveys; two major problems in survey sampling. Our approach addresses both selection- and measurement effects, underlying the possible mode-effects. Furthermore, we allow for not missing at random (NMAR) nonresponse, by considering the nonresponse as another mode. Unlike other approaches proposed in the literature, we do not assume that one of the modes provides unbiased predictors. The existence of such a mode is not guaranteed, and even if it exists, it is not clear how to determine which one it is. The approach is model-based but we cannot think of a proper design-based approach that can deal simultaneously with selection- and measurement effects and NMAR nonresponse, without very strong and generally untestable assumptions. In this article we restricted to binary outcome variables (fitting logistic models in the empirical illustrations), but the proposed approach can be extended to continuous outcomes, with proper modifications.

We propose simple test procedures for testing our model, and in particular, for testing the existence of measurement effects, which are seen to work well in the empirical studies, although more powerful tests can, and should be developed. When applied to proxy surveys, an interesting open question is how to define the different modes. In our empirical study we defined them in an “ad-hoc” manner, but a more founded methodology should be established.

One possible way is to start with as many as possible modes, estimate the means or other characteristics of interest for each mode, and then collapse modes based on proper statistical analysis, so as to stabilize the final results.

The empirical results with the simulated and real data sets are promising and we encourage other researchers to test the approach with their data. We mention again that the approach is applicable in principle also to nonprobability samples, which become more and more popular in recent years with the availability of new “big data” sets.

10. APPENDIX

Let the variables Y,M,X,Z under Model A (y,Y,M,X,Z under Model B) be defined on the probability space $(\Omega,\mathbb{F},\mathrm{P})$. Condition A1(B1) implies the following condition:

(*) the sequences $\{\ell_i(\delta)\}_{i\in S}$, $\{\tilde{\ell}_i(\delta)\}_{i\in S}$ and their first and second derivatives are iid and bounded almost surely respectively.

Denote, $\ell(\delta) = I(M < \overleftrightarrow{M}) log[f(\delta)^Y g(\delta)^{1-Y}] + I(M = \overleftrightarrow{M}) log[f(\delta)+g(\delta)]$, and $\ell_0 = \ell(\delta_0)$. Also, let $\tilde{\ell}(\theta) = I(M < \overleftrightarrow{M})\big(y\log[p_1(\gamma)f(\delta)+(1-p_0(\gamma))g(\delta)]$ $+(1-y)\log[p_0(\gamma)g(\delta)+(1-p_1(\gamma))f(\delta)]\big) + I(M = \overleftrightarrow{M})\log[f(\delta)+g(\delta)]$,

where $p_j(\gamma) = \Pr(D=1 \mid Y=j,W,M;\gamma),\ D = I(Y=y)$ for $M \neq \overleftrightarrow{M},\ j=0,1$.

Proof of theorem 1. (i) First, we show that the model is identifiable, that is, $\ell(\delta)=\ell(\delta^*)$ almost surely implies $\delta=\delta^*$. Let $F_1=\{\omega\in\Omega \mid M(\omega)\neq\overleftrightarrow{M}\}$ and $F_2=\{\omega\in\Omega \mid M(\omega)=\overleftrightarrow{M}\}$. Note that under the condition A1 both sets are non-null. Let $\omega\in F_1$, and suppose that $\delta\neq\delta^*$ but $\ell(\delta)=\ell(\delta^*)$. Suppose first that $\alpha\neq\alpha^*$.

(1) . $H_1(\alpha)H_2(\beta)=H_1(\alpha^*)H_2(\beta^*) \Rightarrow \dfrac{H_1(\alpha)}{H_1(\alpha^*)}=\dfrac{H_2(\beta^*)}{H_2(\beta)}$

Under Condition A2, there exists a variable X_v not included in Z. Let, $H_1^v(\alpha) = \partial H_1(\alpha)/\partial X_v$ and denote by X_{-v} the vector of covariates in X excluding X_v. Taking the partial derivative of (1) with respect to,yields X_v

$$\frac{H_1^v(\alpha)H_1(\alpha^*) - H_1^v(\alpha^*)H_1(\alpha)}{(H_1(\alpha^*))^2} = 0 \Rightarrow \frac{\partial \log[H_1(\alpha)H_1^{-1}(\alpha^*)]}{\partial X_v} = 0 \tag{2}$$

By integrating (2) with respect to X_v and using the notation in (A3),

$$\log H_1(\alpha) - \log H_1(\alpha^*) = h_1(X, \alpha, \alpha^*) = \psi(X_{-v}), \tag{3}$$

where $\psi(X_{-v})$ is some differentiable function with respect X_{-v}. Taking the partial derivative of (3) with respect to X_v implies $\partial h_1(X)/\partial X_v = 0$, which contradicts A2 and hence $\alpha = \alpha^*$. It remains to show that $\beta = \beta^*$. Substituting $\alpha = \alpha^*$ in (1), it follows from A1 that $\beta = \beta^*$ and thus $\delta = \delta^*$.

A similar proof applies when considering $\beta \neq \beta^*$ and when there exists a variable Z_v in Z not included in X.

Consider now $\omega \in F_2$. Using similar arguments to above, we can show that $f_{\tilde{M}}(\delta)$ and $g_{\tilde{M}}(\delta)$ are each identifiable and by condition A3, they are linearly independent. (The functions $f_M(\delta)$ and $g_M(\delta)$ are defined in Section 3.1). Hence, $f_{\tilde{M}}(\delta) + g_{\tilde{M}}(\delta) = f_{\tilde{M}}(\delta^*) + g_{\tilde{M}}(\delta^*) \Rightarrow \delta = \delta^*$. This completes the proof of identifiability.

Second, the compactness of the parameter set and the identifiability property implies by the information inequality that $E\ell(\delta_0) - \max_{\delta \in \Delta} E\ell(\delta) \geq 0$.

Third, by (*) and Theorem A.2.2 in White (1994),

$$\max_{\delta \in \Delta} | L_n(\delta) - E\ell(\delta) | \rightarrow_{a.s} 0. \tag{4}$$

Given the results so far and using similar arguments to those used in the proof of Theorem 3.4 of White (1994), it follows that $\hat{\delta}_n \rightarrow_{a.s.} \delta_0$, thus completing the first part of the theorem.

(ii) Note that $E(\nabla_{\delta_0} \ell_0) = \nabla_{\delta_0}(E\ell_0) = 0$. The left hand side equality follows by Condition (*). The right hand side equality follows from Condition A4. The identifiability of the model shown in part (i) and Theorem 1 of Rothenberg (1971) implies that for sufficiently large n, C_{0n} is positive definite. The last two results imply by the Lindberg-Levy Central Limit Theorem,

$$\sqrt{n}C_{0n}^{-1/2}\nabla L_n(\delta_0) \to_D N(0, I_k) . \tag{5}$$

Further, by (*), Condition A4 and Theorem A.2.2 in White (1994),

$$\max_{\delta\in\Delta} \| D_n(\delta) - E(D_n(\delta)) \| \to_{a.s} 0 . \tag{6}$$

Thus, by the strong consistency of $\hat{\delta}_n$ shown in the first part,

$$\| D_n(\hat{\delta}_n) - D_n(\delta_0) \| \to_{a.s} 0 . \tag{7}$$

By (5)-(7), and Theorem of 6.2 in White (1994),

$C_n^{-1/2} D_n \sqrt{n}(\hat{\delta}_n - \delta_0) \to_D N(0, I_k)$. Q.E.D

Proof of Theorem 2. We again start by proving the likelihood identifiability (Equation 3.6). Denote as before, $p_j = h_3^j(\gamma) = \Pr(D=1 \mid Y=j, W, M; \gamma)$ for $M_i \neq \overleftrightarrow{M}$, $j = 0,1$ and $p = h_3(\gamma) = (h_3^0(\gamma), h_3^1(\gamma))'$. Using the same steps as in the proof of theorem 1, it can be shown that under Condition B1, the probabilities $f_M(\delta)$ and $g_M(\delta)$ are both identifiable. Hence, by Condition B1, for the set $\tilde{F}_2 = \{\omega \in \Omega \mid M(\omega) = \overleftrightarrow{M}\}$, $\tilde{\ell}(\theta) = \tilde{\ell}(\theta^*) \Rightarrow \theta = \theta^*$, similarly to the first part of Theorem 1. For $\omega \in \tilde{F}_1 = \{\omega \in \Omega \mid M(\omega) \neq \overleftrightarrow{M},\ y(\omega) = 1\}$, the contribution to the likelihood is given by,

$$G(\delta, p) = p_1 f_M(\delta) + (1 - p_0) g_M(\delta) \tag{8}$$

(compare with 3.6). Thus, under Condition B1, $G(\delta, p)$ is identifiable in the sense that,

$$G(\delta, p) = G(\delta^*, p^*) \Rightarrow (\delta, p) = (\delta^*, p^*) \ , \tag{9}$$

and since $h_3(\gamma)$ is identifiable, we have that if $p = p^* \Rightarrow \gamma = \gamma^*$ and $\theta = \theta^*$. By repeating the same arguments as above, we establish the identifiability of the model also for the set $\omega \in \tilde{F}_3 = \{\omega \in \Omega \mid M(\omega) \neq \overleftrightarrow{M},\ y(\omega) = 0\}$. The rest of the proof of strong consistency (part i) and asymptotic normality (part ii) of the MLE, is similar to the proof of Theorem 1.

11. REFERENCES

Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis. *The Annals of Statistics,* **45**, 77-120.

Berndt, E.K., Hall, B.H., Hall, R.E. and Hausman, J.A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3**, 653–66.

Biemer, P.P. (1988). Measuring data quality. In: *Telephone Surveys Methodology*. New York: Wiley and Sons.

Biemer, P.P. (2001). Nonresponse Bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, **17**, 295-320.

Dempster, A.P., Laird, N. M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1–38.

De Leeuw, E.D. (2018). Mixed-mode: past, present and future. *Survey Research Methods*, **12**, 75-89.

De Leeuw, E.D., Suzer-Gurtekin, Z. and Hox, J. (2018). The design and implementation of mixed mode surveys. In: *Advances in Comparative Survey Methodology*. New York: Wiley and Sons.

Dillman, D. (2000). Mail and internet surveys. New York: Wiley and Sons.

Hox, J., de Leeuw, E.D., and Klausch, T. (2017). Mixed mode research: issues in design and analysis. In: *Total Survey Error in Practice.* Wiley Series in Survey Methodology.

Dillman, D. A. and Christian, L. M. (2003). Survey mode as a source of instability in responsesacross surveys. *Field Methods*, **15,** 1–22.

Follmann, D.A. and Lambert, D. (1991). Identifiability of finite mixture of logistic regression models. *Journal of Statistical Planning and Inference,* **27**, 375-381.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken NJ: Wiley and Sons.

Hosmer, D.W. and Lemeshow, S. (1980). A goodness-of-fit test for multiple logistic regression model. *Communications in Statistics*, A, **10**, 1043-1069.

Hosmer, D.W. Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley and Sons.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association***, 77,** 89-96.

Kalsbeek, W.D. and Agans, R.P. (2007). Sampling and weighting in household telephone surveys. In: *Advances in Telephone Survey Methodology*. New York: Wiley and Sons.

Kolenikov, S. and Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology*, **2**, 126-158.

Kormendi, E. (1988). The quality of income information in telephone and face to face surveys. In *Telephone Survey Methodology*, New-York: John Wiley and Sons.

Moore, J.C. (1988). Self/Proxy response status and survey response quality: A review of literature. *Journal of Official Statistics*, **4**, 155-172.

O'Muircheartaigh, C. (1991). Simple response variance: estimation and determinants. *Measurement Errors in Surveys*, New York, Wiley.

Park, S., and Kim, J.K. and Park, S. (2016). An imputation approach for handling mixed mode surveys. *Annals of Applied Statistics*, **10**, 1063-1085.

Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179-189.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 317-337.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, **3**, 425–483.

Pfeffermann, D. (2017). Bayes-based Non-Bayesian Inference on Finite Populations from Non-representative Samples**.** A Unified Approach. *Calcutta Statistical Association (CSA) Bulletin*, **69**, 35-63.

Pfeffermann, D. and Landsman, A. (2011). Are private schools really better than public schools? Assessment by methods for observational studies. *Annals of Applied Statistics*, **5**, 1726-1751.

Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under non ignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, **27**, 181-209.

Pfeffermann , D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, **61**, 166-186.

Rothenberg, T.J. (1971). Identification in parametric models. *Econometrica*, **39**, 577-591.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-590.

Rubin, D. B. *(1987). Multiple Imputation for Nonresponse in Surveys*. New York: Wiley and Sons.

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

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, **79**, 516-524.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychlogical Bulletin*, **133**, 859-883.

Turner, C.F., Ku, L., Rogers, S., Lindberg, L., Pleck, J. and Sonenstein F. (1998). Adolescent sexual behavior, drug use and violence: increased reporting with computer survey technology. *Science*, **280**, 867-873.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, **74**, 1027-1045.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2014). Evaluating mode effects in mixed-mode survey data using covariate adjustment models. *Journal of Official Statistics*, **30**, 1-24.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

White, H. (1994). *Estimation*, *Inference and Specification Analysis*. Cambridge University Press.

Wilson, P. (2015). The Misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, **127**, 51-53.