Proceedings of Q2006
European Conference on Quality in Survey Statistics

A Model for Statistical Inference based on Mixed Mode Interviewing

Fannie Cobben, Barry Schouten, and Jelke Bethlehem[1]

***Abstract**: Household surveys can be conducted using various data collection modes. Each individual data collection mode has its shortcomings. Face-to-face interviewing is expensive. Not every household has a telephone/Internet connection and can be approached by CATI resp. a Web survey. Mail surveys have a low response rate. Mixing data collection modes provides an opportunity to compensate for the weakness of each individual mode. This can reduce costs and at the same time increase the response. It is even possible to reduce the selectivity of the response beforehand. For this purpose, sampled persons or households can be allocated to a specific mode based on known background characteristics.*

An optimal mixed mode strategy may still be in the future, but suppose that we have a survey administration system with decision rules that can support any strategy of mixed mode data collection. Such a system gives us a number of datasets, collected through different modes, for the same survey. How can we combine this data, so that we can use it for statistical inference?

We extend the sample selection model (see Heckman (1979)) so that it can be used to aggregate data from general mixed mode strategies and at the same time adjust for non-response bias.

***Keywords**: Non-response bias, mixed mode data collection, sample selection model*

1. Introduction

1.1 Mixed mode data collection

In this paper, we assume that the data collection covers the entire population. We restrict ourselves to data collection during the response phase, i.e. we disregard the contact phase (e.g. notification letters, screener calls). We also do not regard the mixed mode variant where the choice of data collection mode is left to the respondent. De Leeuw (2005) describes two different mixed mode data collection designs. The first design is a concurrent system. The sample is divided in groups that are approached by

[1] Fannie Cobben, Statistics Netherlands, Methods and Informatics Department, room 241A, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, e-mail: fcbn@cbs.nl; Barry Schouten, Statistics Netherlands, Methods and Informatics Department, P.O. Box 4000, 2270 JM Voorburg, The Netherlands; Jelke Bethlehem, Statistics Netherlands, Methods and Informatics Department, P.O. Box 4000, 2270 JM Voorburg, The Netherlands

a different mode, but at the same time. See Figure 1. The other design is a sequential design. All sample elements are approached by one mode. The non-respondents are then followed up by a different mode than used in the first approach. This process can be repeated, see Figure 2.

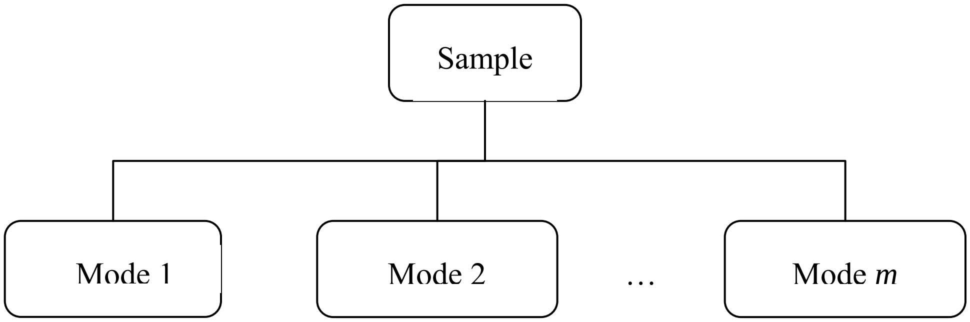

Figure 1: Concurrent mixed mode design.

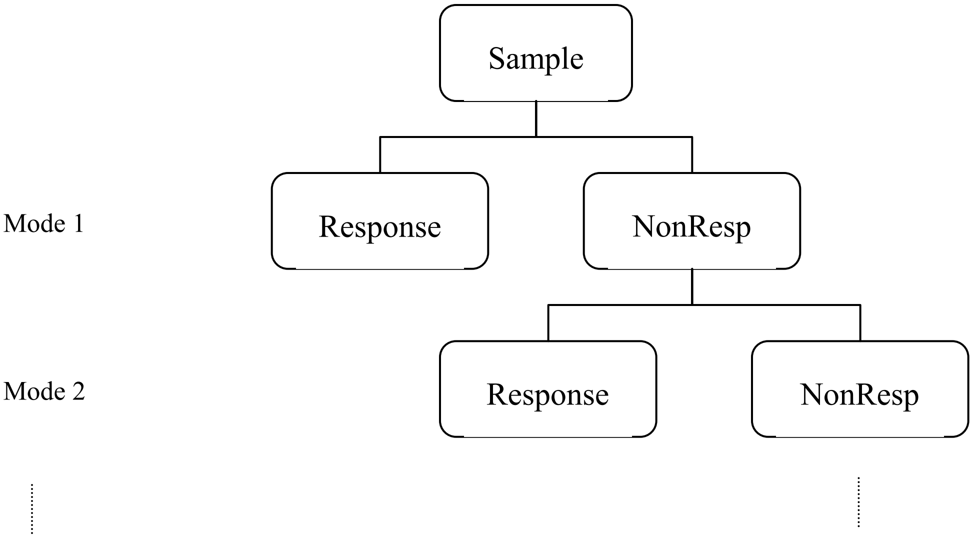

Figure 2: Sequential mixed mode design.

Modes differ in various aspects, e.g. timeliness or costs. Biemer and Lyberg (2003) discuss optimal designs for unimode data collection. See Pierzchala (2006) for an overview of the differences in cognition and response. Because of these disparities, there are mode effects. A mode effect occurs if the answers of a respondent differ when asked the same question in a different mode. It is difficult to evaluate mode effects. In this paper, we make some assumptions regarding these effects. First, we assume that there are no questionnaire effects. As De Leeuw (2005) notes, the questionnaires need to be equivalent in a cognitive way (and can thus vary by mode without causing a mode effect). In fact, we do not include measurement errors in our models in general.

1.2 Response process

Before a person actually participates in a survey, there are some hurdles to be taken. First, contact has to be made before a person can decide to comply with the survey request or not. When contact has been made, the person must be able to participate. A person can be unable to participate due to language problems, or may be able to speak the language but is unable to cooperate due to a longtime illness. When contact has been made, and a person is also able to participate, the last hurdle is the willingness to comply with the request.

The response process can be decomposed in a number of stages, see Figure 3. These decompositions differ between modes. One important distinction is the assistance of an interviewer. Especially the last step, i.e. refusal or cooperation, is influenced by the interaction between the interviewer and the sample person. See Groves *et al.* (1992). But there are other distinctions as well. For instance, in a Web as well as a paper survey it is only observed whether a person participates in the survey or not. The actual reason for the non-response is unclear. It could be a non-contact, a language problem, a longtime illness as well as a refusal.

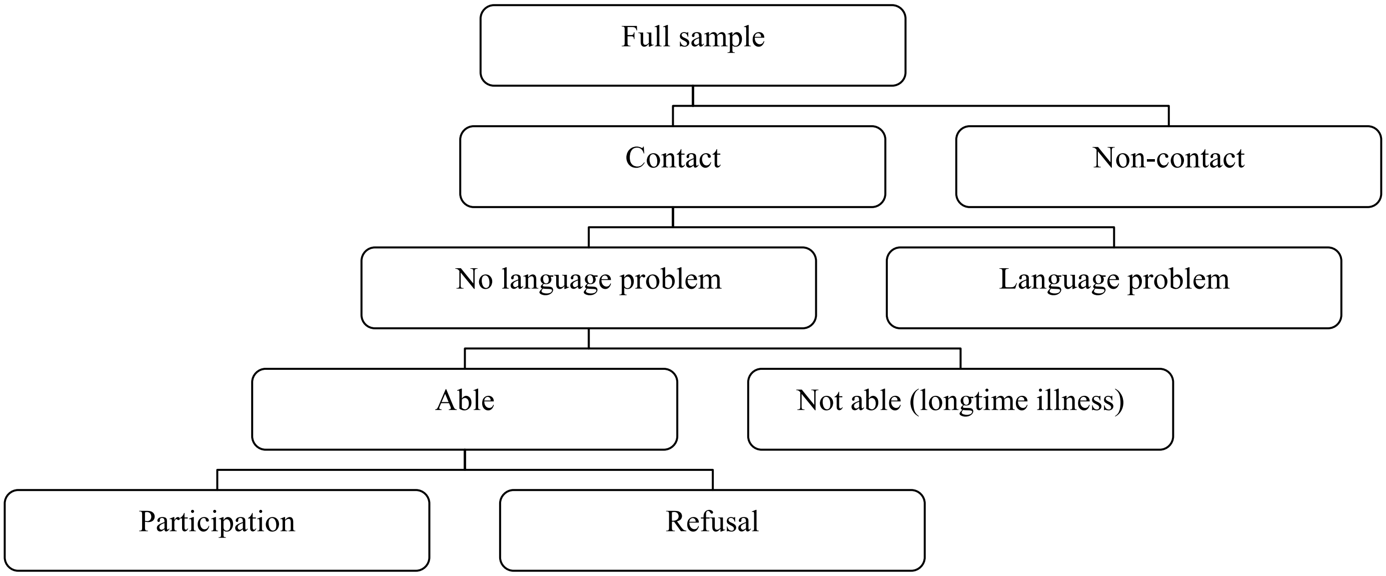

Figure 3: The response process

Each of these steps relates to different characteristics of the respondent and the data collection. E.g. the non-respondents due to language problems will have a different profile than the refusers. By distinguishing between these types of non-response and the incorporation of instrumental variables we can better explain the process and eventually better adjust for non-response bias. In the literature, this approach is suggested as well, see e.g. Lepkowski and Couper (2002) or Nicoletti and Peracchi (2005).

1.3 Outline

The aim of our research is to develop a model that combines data collected by mixed modes (both concurrent and sequential), thereby accounting for the different response processes in each of the individual modes. We translate and extend the model by Heckman (1979) to mixed mode and response models. We follow the bivariate probit model like in Van de Ven and Van Praag (1981).
In section 2, we present models to combine data in both the concurrent and the sequential mixed mode design. Section 3 discusses a model for the response process as described in subsection 1.2. A combination of the designs and the model for the response process is outlined in section 4, as well as an agenda for future research.

2. Mixed mode models

2.1 Concurrent

Recall that the concurrent mixed mode design assigns sample persons to a specific mode. All sample persons are thus approached at the same time but in different modes. One can think about optimal allocation strategies that reduce non-response bias and increase response. We assume that the allocation probabilities of the sample persons to a specific mode are known beforehand. The allocation probability for sample person *i* is denoted by $\eta(x_i)$. We do not include the entire response process yet but only distinguish between response (R) and non-response (NR). In every mode *m* there is an underlying participation decision that determines the response probability $\rho^m(x_i)$. This participation decision in mode *m* is denoted by I_m. The model for the concurrent mixed mode with two modes looks like Figure 4.

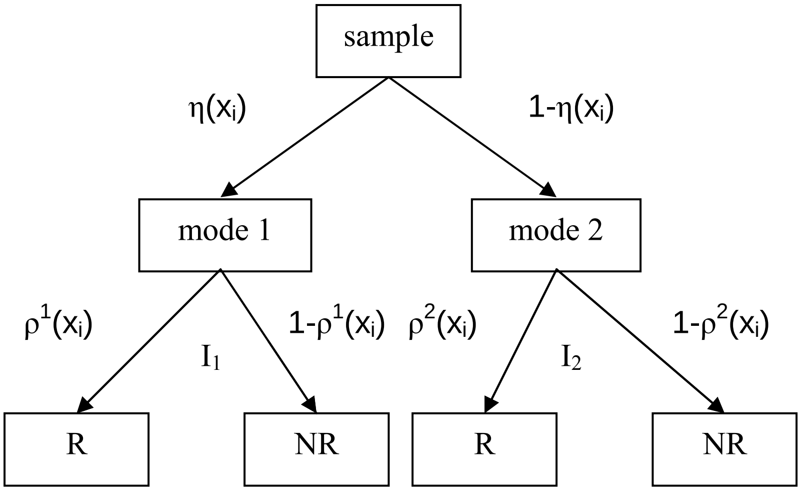

Figure 4: Concurrent mixed mode model.

The answer to a survey question is obtained when the sample person is assigned to mode 1 and participates in mode 1 or when he/she is allocated to mode 2 and participates in this mode. The model can be described in a two-step manner. Conditional on the mode allocation and the response process for the person in the assigned mode, an answer to the survey question is obtained. This can be described in a similar way as the sample selection model proposed by Heckman (1979). Let us first introduce some notation. Let the target population of a sample survey consist of *N* individuals 1, 2,…, *N*. Let *Y* denote a target variable of the survey. Associated with each individual *k* is a value Y_k of this target variable. Assume that the aim of the sample survey is to estimate the population mean of the target variable

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \; . \qquad (2.1.1)$$

Furthermore, let *X* be a vector of auxiliary variables or covariates, with values X_k, for *k* = 1, 2, …, *N*. The sample selection model consists of two stages. In the context of survey participation the first stage models the participation of a person in the survey. Consequently, in the second stage the outcome to the survey is estimated while making use of the information from the first stage by correcting for the persons that did not participate. With the concurrent mixed mode design this process occurs for groups of persons in different modes; determined by the allocation probabilities $\eta(x)$. There is a latent variable I^* that determines the participation. However, this variable is not observed. We only observe the outcome of the process (a response or a non-response in this case). In equation (2.1.2) the model for the first stage is described.

$$Mode_i = \begin{cases} 1, with\ probability\ \eta(x_i) \\ 2, with\ probability\ 1-\eta(x_i) \end{cases}$$

$$\begin{aligned} & For\ \ m = 1,2: \\ & I^*_{i,m} = \beta_i^m X_i^m + \delta_i^m \\ & I_{i,m} = I\{I^*_{i,m} \geq 0\} \end{aligned} \tag{2.1.2}$$

The parameters β_i^m, δ_i^m are resp. the vector of coefficients and the random error term for mode *m*. We assume that each person has an answer to the survey and thus a value for the target variable. We just do not always observe it. This can also be modelled as a latent variable equation, where the target variable *Y* is the latent variable that can be explained by auxiliary variables *X* and a certain random error *μ*. See equation (2.1.3). This variable is observed conditional on the outcome of the participation process in the second part of equation (2.1.2).

$$\begin{aligned} & Y_i^* = \gamma X_i + \mu_i \\ & Y_i = \begin{cases} Y_i^* \ \ if\ \ \text{mod}e\ 1\ and\ \ I_1 = 1\ or\ \ \text{mod}e\ 2\ and\ \ I_2 = 1 \\ - \ \ else \end{cases} \end{aligned} \tag{2.1.3}$$

Equation (2.1.2) and (2.1.3) are linked by their error terms δ^1, δ^2, μ. This is expressed in the correlation structure. Usually a multivariate normal distribution is assumed. See equation (2.1.4).

$$\begin{pmatrix} \delta^1 \\ \delta^2 \\ \mu \end{pmatrix} \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & \sigma^2 \end{pmatrix} \right) \tag{2.1.4}$$

This model can e.g. be estimated by maximum likelihood, where the likelihood equals

$$\begin{aligned} L = \sum_{i=1}^{n} \Big[& \eta(x_i)\{I_{i,1}P(Y_i = y_i; I_{i,1} = 1 \mid X_i) + (1 - I_{i,1})P(Y_i = y_i; I_{i,1} = 0 \mid X_i)\} \\ & + (1-\eta(x_i))\{I_{i,2}P(Y_i = y_i; I_{i,2} = 1 \mid X_i) + (1 - I_{i,2})P(Y_i = y_i; I_{i,2} = 0 \mid X_i)\} \Big] \end{aligned} \tag{2.1.5}$$

2.2 Sequential

In a sequential mixed mode design, the entire sample is first approached by one specific mode. The non-respondents to that mode are then followed-up by another mode. This process can be repeated. Like in the concurrent design, there is a latent variable that determines the participation. The difference is that the process is now repeated in time, for different modes and only for the non-respondents in earlier modes. See Figure 5.

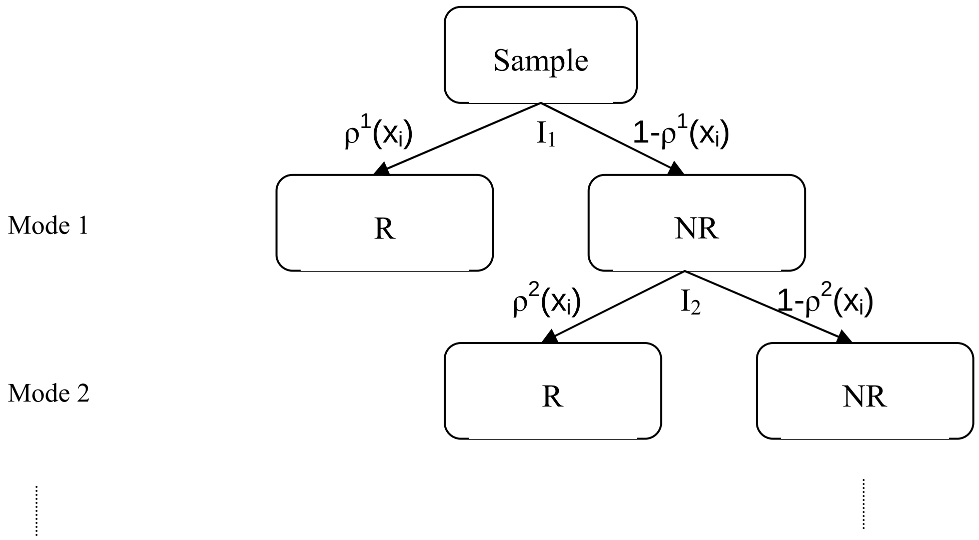

Figure 5: Sequential mixed mode model.

In the same line of reasoning as the concurrent model, we can describe this model as follows in equations (2.2.1) – (2.2.4).

$$
\begin{aligned}
I_{i,1}^{*} &= \alpha X_i + \varepsilon_i \qquad I_{i,1} = I\{I_{i,1}^{*} \geq 0\} \\
I_{i,2}^{*} &= \beta X_i + \delta_i \qquad I_{i,2} = I\{I_{i,2}^{*} \geq 0\}
\end{aligned}
\tag{2.2.1}
$$

$$
\begin{aligned}
& Y_i^{*} = \gamma X_i + \mu_i \\
& Y_i = \begin{cases} Y_i^{*} & if\ I_{i,1} = 1\ or\ I_{i,2} = 1 \\ - & else \end{cases}
\end{aligned}
\tag{2.2.2}
$$

$$
\begin{pmatrix} \varepsilon_i \\ \delta_i \\ \mu_i \end{pmatrix} \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & \sigma \end{pmatrix} \right)
\tag{2.2.3}
$$

$$
\begin{aligned}
L = \sum_{i=1}^{n} \big[& I_{i,1} P(Y_i = y_i; I_{i,1} = 1 \mid X_i) + (1 - I_{i,1}) I_{i,2} P(Y_i = y_i; I_{i,1} = 0; I_{i,2} = 1 \mid X_i) \\
& + (1 - I_{i,1})(1 - I_{i,2}) P(Y_i = y_i; I_{i,1} = 1; I_{i,2} = 0 \mid X_i) \big]
\end{aligned}
\tag{2.2.4}
$$

The main difference between the concurrent and the sequential mixed mode model is in the correlation structure of the error terms and, consequently, in the likelihood. It becomes clear in the distribution of the error terms from equations (2.2.1) and (2.2.2) that now the process is sequential. For a person to be a respondent in mode *m*, *m > 1*, this person has had to be a non-respondent in all the modes < *m*. This condition is translated by the correlation between the participation processes whereas in the concurrent mixed mode model these equations have a zero correlation, see equation (2.1.4).

3. Response model

In subsection 1.2 we describe the response process. We ‘peel off’ the entire process to distinguish between different types of non-response. Figure 3 gives a graphical display. As we already motivated, it is important to make a distinction between these groups because they can be very different. If, and only if, there are instrumental variables available that partially explain the various stages, we believe it to be beneficiary to model them separately. It can be used to better estimate the survey outcomes, i.e. to better adjust for non-response bias.

The model that is described in this section is able to combine all sources of non-response and to estimate a survey outcome incorporating instrumental information. Without this information one does not need to model stages separately from a non-response adjustment perspective. However, if one is interested in the separate stages one could leave the distinction. In that case the equations can be estimated separately as the correlations are introduced by the incorporation of instrumental variables alone. This is an important feature of the response model, because by combining the different non-response types into one model, the mass of the observations is preserved which means that the model has more explanatory power. See Figure 6 for a graphical representation. Of course, other decompositions can be modelled analogously.

Nicoletti and Peracchi (2005) make a similar distinction. They distinguish non-contact and refusal as causes of non-response. They do, however, not model the answers to the survey questions and focus solely on the characteristics of these two types of non-response.

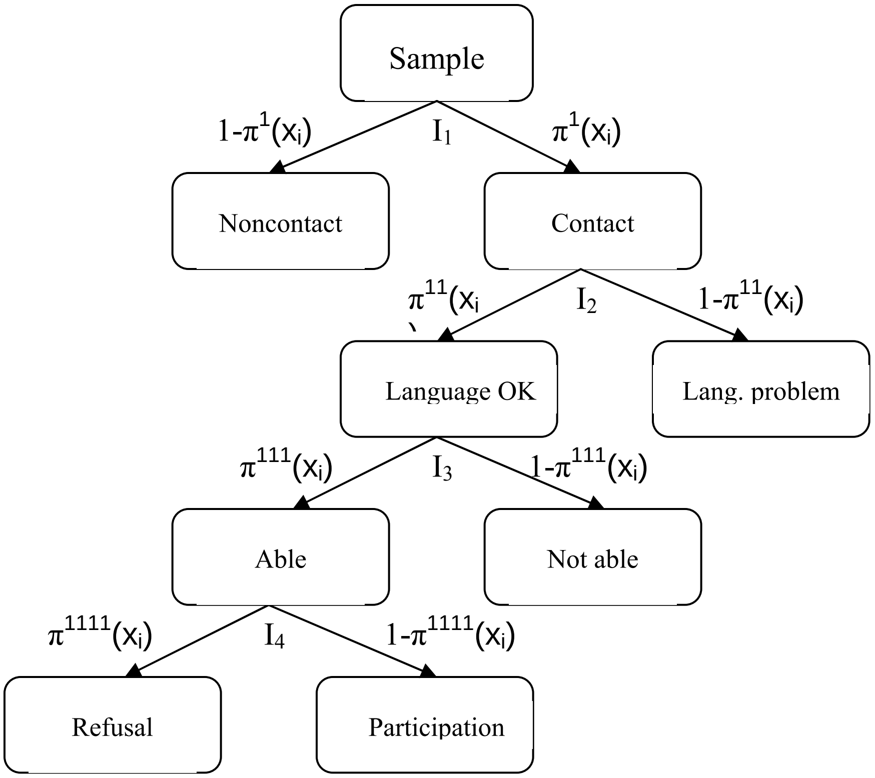

Figure 6: The response model.

The model is again similar to the models in section 2.1 and 2.2, in that respect that we use a multiple stage structure as in the concurrent mixed model and a dependent structure as in the sequential mixed mode model.

There are *j = 4* selection equations that determine whether a person passes through to a next stage in the response process, which ends with refusal or participation. Hence there are 5 possible outcomes to the response process: *Non-contact, language problem, not able due to long-time illness, refusal* or *participation*. Equation (3.1) displays the selection equations for the *j = 4* stages in the response process. The final probability of response (π) can be described by the probabilities determined by the selection equations ($\pi^{1}, \pi^{11}, \pi^{111}, \pi^{1111}$). See equation (3.2).

$$\begin{aligned} I_{ij}^{*} &= \beta_j X_{ij} + \varepsilon_{ij} \\ I_{ij} &= I\left\{ I_{ij}^{*} > 0 \right\} \end{aligned} \tag{3.1}$$

$$\begin{aligned}
&\pi_1 = P(I_1 = 1) \\
&\pi_{11} = P(I_2 = 1 \mid I_1 = 1) \\
&\pi_{111} = P(I_3 = 1 \mid I_2 = 1; I_1 = 1) \\
&\pi_{1111} = P(I_4 = 1 \mid I_3 = 1; I_2 = 1; I_1 = 1) \\
&\pi = P(respons) = \pi_1 * \pi_{11} * \pi_{111} * \pi_{1111}
\end{aligned} \tag{3.2}$$

The outcome for the target variable is only observed when a person participates in the survey. This is modelled in (3.3).

$$\begin{aligned}
&Y_i^* = \gamma X_i + \mu_i \\
&Y_i = \begin{cases} Y_i^* \ \ iff\ I_{i4} = 1 \\ - \ \ else \end{cases}
\end{aligned} \tag{3.3}$$

The distribution of the error terms is again assumed to be multivariate normal, see equation (3.4).

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \mu_i \end{pmatrix} \sim N_5 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{21} & 1 & 0 & \rho_{24} & \rho_{25} \\ \rho_{31} & 0 & 1 & \rho_{34} & \rho_{35} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & \sigma \end{pmatrix} \right) \tag{3.4}$$

Only those auxiliary variables are used in equations (3.1) that relate to the corresponding causes of non-response. This enables us to use the information of interest exactly there where it adds explanatory power. For instance, information about the interviewer can be included when explaining participation or refusal but has nothing to do with a person being longtime ill or not.

Again we can estimate this model by maximum likelihood, see (3.5). When these models are estimated, the estimated parameters are used to estimate the mean of the target variable (2.1.1).

$$\begin{aligned}
L = \sum_{i=1}^{n} \Bigg[& \prod_{j=1}^{4} I_{ij} P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1; I_{i4} = 1 \mid X_i) + \\
& + I_{i1} I_{i2} I_{i3} (1 - I_{i4}) P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1; I_{i4} = 0 \mid X_i) \\
& + I_{i1} I_{i2} I_{i3} P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1 \mid X_i) + \ldots \Bigg]
\end{aligned} \tag{3.5}$$

4. Future research: combination of mixed mode- and response models

In section 2 we present two models for a mixed mode strategy, a *concurrent*- and a *sequential* model. Additionally, in section 3 a response model to adjust for non-response bias is discussed. This response model makes a clear distinction between different causes of non-response and uses instrumental variables to explain the stages in the response process.

The mixed mode models and the response model are the basic ingredients for a general framework to combine data from different mixed mode designs, thereby accounting for the different response processes of these modes and making optimal use of all available information. Response models can simply be inserted into the mixed mode models wherever necessary or needed. In the concurrent design, this means response models are inserted at the leaves of the tree. In the sequential design it is somewhat more complicated. Extending the model with different types of non-response would imply that all sources of non-response proceed to a next mode. In general this is not true. One may for instance follow up non-contacts only. Insertion there depends on the follow up strategy chosen.

There are a number of important issues that need further research. First, the estimation procedure of the models needs to be worked out. Second, we need to find a strategy to select variables. Third, we need to extend the models to designs with unequal inclusion probabilities.

Estimation by maximum likelihood, as suggested in sections 2 and 3, is an option but the models can become very large, hence leading to high dimensional parameter spaces. Furthermore, the likelihoods cannot be written in closed form so that we need to resort to numerical methods making the estimation complex and burdensome. A possible alternative is Bayesian estimation of the models, see Albert and Chib (1993) and Groenewald and Mokgatlhe (2004).

The number of auxiliary variables is in general too large to enter them all in the models simultaneously. We, therefore, need a procedure to enter variables to each of the equations. Since there can be many such equations in a complicated mixed mode design, this is not straightforward.

In many cases samples are stratified based on the expected variances within strata. We cannot simply add the inverse inclusion probabilities as weights to the equations. Future research will be directed at efficient estimation schemes, strategies for the selection of variables, and at general sampling designs.

References

Albert, J.H. and Chib, S. (1993), “Bayesian analysis of binary and polychotomous response data”, *Journal of the American Statistical Association,* Vol. 88, No. 422, p. 669 - 679

Biemer, P.P. and Lyberg, L.E. (2003), *Introduction to Survey Quality,* Wiley series in survey methodology

De Leeuw, E.D. (2005), 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics,* Vol. 21, No. 2, pp. 233 – 255

Groenewald, P.C.M. and Mokgatlhe, L. (2004), “Bayesian computation for logistic regression”, *Computational Statistics and Data Analysis*, Vol. 48, p. 857 – 868

Groves, R.M. and Couper, M.P. (1998), *Nonresponse in Household Interview Surveys,* Wiley series in probability and statistics: survey methodology section

Groves, R.M., Cialdini, R.B. and Couper, M.P. (1992), “Understanding the decision to participate in a survey”, *Public Opinion Quarterly*, Vol. 56, No. 4, p.475 – 495

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica*, Vol. 47, No. 1, pp. 153 – 161

Lepkowski, J.M. and Couper, M.P. (2002), “Nonresponse in the Second Wave of Longitudinal Household Surveys”, In R.M. Groves et al. (eds.) *Survey Nonresponse*, New York, Wiley, p. 259 – 272

Nicoletti, C. and Peracchi, F. (2005), “Survey response and survey characteristics: microlevel evidence from the European Community Household Panel”, *Journal of the Royal Statistical Association Series A*, 168, Part 4, p. 763 – 781

Pierzchala, M. (2006), “Disparate Modes and Their Effect on Instrument Design”, *Proceedings of the 10th International Blaise Users Conference*, p. 199-209

Van de Ven, W. and Van Praag, B.M.S. (1981), “The demand for deductibles in private health insurance: a probit model with sample selection”, *Journal of Econometrics*, 17, p. 229 – 252