Please do not cite without permission of authors

Using prior wave information and paradata: Can they help to predict response outcomes and call sequence length in a longitudinal study?

Gabriele B. Durrant[1], Olga Maslovskaya[2] and Peter W.F. Smith[3]

[1] Department of Social Statistics and Demography and
ESRC National Centre for Research Methods
School of Social Sciences
University of Southampton, UK
(g.durrant@southampton.ac.uk)

[2] ESRC National Centre for Research Methods (NCRM)
School of Social Sciences
University of Southampton, UK
(om206@soton.ac.uk)

[3] Department of Social Statistics and Demography and
ESRC Administrative Data Research Centre for England
School of Social Sciences
University of Southampton, UK
(p.w.smith@soton.ac.uk)

Address for correspondence:
Gabriele Durrant
Department of Social Statistics and Demography
School of Social Sciences
University of Southampton
SO17 1BJ, Southampton
g.durrant@southampton.ac.uk

Abstract:

In recent years the use of paradata for nonresponse investigations has risen significantly. One key question is how useful paradata, including call record data and interviewer observations, from the current and previous waves of a longitudinal study, as well as previous wave survey information, are in predicting response outcomes in a longitudinal context. This paper aims to address this question. Final response outcome and sequence length (the number of calls/visits to a household) are modelled both separately and jointly for a longitudinal study. Being able to predict length of call sequence and response can help to improve both adaptive and responsive survey designs and to increase efficiency and effectiveness of call scheduling. The paper also identifies the impact of different methodological specifications of the models, for example different specifications of the response outcomes. Latent class analysis is used as one of the approaches to summarise call outcomes in sequences. To assess and compare the models in their ability to predict, indicators derived from classification tables, ROC (Receiver Operating Curves), discrimination and prediction are proposed in addition to the standard approach of using the pseudo R^2 value, which is not a sufficient indicator on its own. The study uses data from Understanding Society, a large-scale longitudinal survey in the UK. The findings indicate that basic models (including geographic, design and survey data from the previous wave), although commonly used in predicting and adjusting for nonresponse, do not predict the response outcome well. Conditioning on previous wave paradata, including call record data, interviewer observation data and indicators of change, improve the fit of the models. A significant improvement can be observed when conditioning on the most recent call outcome, which may indicate that the nonresponse process predominantly depends on the most current circumstances of a sample unit.

Key Words: survey non-response, interviewer call record data, paradata, call sequence, responsive and adaptive survey designs.

Acknowledgement

This work was supported by the UK Economic and Social Research Council (ESRC), 'The Use of Paradata in Cross-Sectional and Longitudinal Research' [grant number: RES-062-23-2997], grant holders: Gabriele B. Durrant, Peter W. F. Smith and Frauke Kreuter, and by the ESRC National Centre for Research Methods, Research Work Package 1, grant number ES/L008351/1.

Data Statement

This study uses wave 1 and wave 2 data from Understanding Society, the United Kingdom Household Longitudinal Study (UKHLS). The data were obtained from the UK Data Archive (http://www.data-archive.ac.uk/). Data reference: University of Essex. Institute for Social and Economic Research and National Centre for Social Research, Understanding Society: Wave 1-3, 2009-2012. 5th Edition. Colchester, Essex: UK Data Archive, November 2013. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-5

1. Introduction

In recent years the use of paradata in survey research has risen significantly (e.g. Groves and Heeringa 2006; Bates *et al*. 2008; Kreuter *et al*. 2010a; Wagner 2013a and 2013b; Durrant *et al*. 2011; Durrant *et al*. 2013a and 2013b, Durrant *et al*. 2015; Potthoff *et al*. 1993; Groves and Couper 1996; Sinibaldi *et al*. 2013; Kreuter 2013; Sinibaldi *et al*. 2014). Paradata may be used for nonresponse investigation and adjustment, measurement error identification and correction, and for the improvement of survey management and design (Kreuter 2013). Several papers have explored the use of paradata for nonresponse adjustment (Kreuter and Kohler 2009; Kreuter *et al*. 2010b; Biemer *et al*. 2013; Hanly 2014; Hanly *et al*. 2015) but concluded that the variables did not contribute much to the enhancement of nonresponse models. Kreuter and Kohler (2009) hypothesised that paradata instead may be more beneficial for the advancement of survey designs, survey processes and data management. Furthermore, the use of paradata for longitudinal surveys is significantly underexplored, although here the greatest benefits may lie, given the rich information about sample cases from previous waves. We are in fact aware of only one conference presentation in this area (Lagorio 2015).

This paper here aims to address this shortcoming and investigates the use of paradata, including call record data and interviewer observation data, from the previous and current wave as well as previous survey information for the prediction of (final) response outcomes in the current or future waves of a longitudinal study. Standard response models have been shown to perform poorly in terms of prediction, usually with a (pseudo) R^2 value of well under 8% (Olson *et al*. 2012; Olson and Groves 2012; West and Groves 2013). This indicates that the response process may be either very difficult to predict given standard variables and methods or that the response process is a more or less random process that is hard to predict by nature. The hope is

that paradata variables and call history information as well as the exploration of different model specifications can lead to improvements in the prediction of response outcomes. Another focus of the paper is on how best to incorporate paradata (from a previous or current wave) into the model. To summarise call record information from the previous wave we propose a latent class analysis approach (Magidson and Vermunt 2004), which to our knowledge has not yet been used for paradata investigations. We also explore the inclusion of derived simple summary measures (e.g. the proportions of different call outcomes, such as proportion of noncontacts). The study uses data from Understanding Society, a large-scale longitudinal survey in the UK, which benefits from the inclusion of rich paradata and information on call records in all waves.

The paper is motivated by earlier exploratory work of interviewer calls (visits) of a face-to-face panel study using sequence analysis, which identified both response outcome and sequence length as important identifying features of call record data (Durrant *et al*. 2016). The call sequence length is defined as the number of calls until final response outcome of a household is reached. The specification of models therefore takes account of both response outcome and sequence length simultaneously. Here in this paper, both phenomena are modelled separately and jointly using logistic and multinomial models respectively. Different model specifications are explored, including various definitions of the dependent variables. The models account for the clustering of households within interviewers by robust standard error estimation. This paper extends previous work which explored models to analyse response outcome of sequences in the case of a cross-sectional survey or for the first wave of a longitudinal survey, where previous wave paradata and previous wave survey variables are not available (see Durrant *et al*. 2015). Here in contrast, prediction of response outcomes in the context of a longitudinal survey taking into account prior wave information is investigated.

The ability of the models to predict response outcomes and sequence length is usually assessed in the nonresponse literature via the pseudo R^2 statistic. However, this indicator was not found sufficient (Plewis *et al*. 2012). We propose additionally the use of a range of indicators derived from classification tables, ROC curves (Receiver Operating Curves) and concepts borrowed from epidemiology such as discrimination and prediction (see also Plewis *et al*. 2012; Agresti 2013).

Unlike some of the previous literature, our analysis does not aim to improve nonresponse adjustment (although this may also be possible in principle) but to enhance survey data management processes. Focussing on both response outcomes and sequence length enables survey researchers to assess the likelihood of a household or groups of households to be successfully contacted and to establish the number of calls it may take to obtain the final response outcome. The aim of this paper is hence to improve both efficiency and effectiveness of interviewer calls. The models can make contributions to both adaptive and responsive survey designs informing improvements in either current or future survey designs respectively. More specifically, for survey researchers it may be of particular relevance to predict long and unsuccessful call outcomes. The ability to identify such cases early on in the call process (either before data collection or after the first, second or third contact attempt) would enable the reduction of survey data collection costs. Knowing that a household will require many calls and is very likely to end with an unsuccessful call outcome will enable survey designers to make informed decisions for the allocation of tailored treatments, such as to stop calling or to increase data collection efforts to alter the likely outcome (e.g. to offer an incentive or send a different interviewer).

Summarising the key research questions to be addressed in this paper, they are:

1. Can predictions of nonresponse models be improved in longitudinal surveys when information from a previous and current wave, including survey data and paradata (i.e. interviewer observation data and call record data) are included?
2. How should the variables best be entered into the models (for example via summary measures or a latent class analysis approach)?
3. Which assessment criteria are best used to compare the ability of nonresponse models to predict the outcome (in addition to the commonly used measure of the pseudo R^2 value)?
4. Can we predict long and unsuccessful call outcomes early on in the data collection process (before data collection or after just one, two or three calls) to improve effectiveness and efficiency of adaptive and responsive survey designs?

The remainder of the paper is structured as follows. Section 2 describes the data and analysis sample. The analysis approach and the methods to assess the different models are described in section 3. Then, results are presented from the separate and joint model specifications. The final section summarises the main findings and discusses implications for survey practice.

2. Data

2.1 Understanding Society – the UK Longitudinal Household Survey

This paper uses data from the first two waves of the UK longitudinal household survey, Understanding Society. The survey has the advantage that it contains rich call record data and a wide range of interviewer observations variables. It is exceptionally large and covers a comprehensive number of variables. Also, only interviewers with a high interviewing

qualification and experience were selected for the survey. The survey has a multi-stage sample design with clustering and stratification, and households are clustered within interviewers. All adult household members (age 16 and older) are asked to respond and the same individuals are re-interviewed in successive waves. Wave 1 data collection took place between January 2009 and March 2011 and wave 2 data collection was conducted between January 2010 and March 2012. For wave 1, interviewers make personal visits to households with interviews carried out using computer assisted personal interviewing (CAPI). In wave 2, households again receive face-to-face interviews (apart from some households that used to be part of the BHPS sample. However, these cases only joined the survey at a later stage and are therefore not of relevance here, see section 2.2). The interviewing protocol requires a minimum of six calls to be made at each sampled address before it is considered unproductive, but interviewers are encouraged to make further calls where possible (McFall 2012). At the beginning of each wave, i.e. at the time of the first call, interviewers collect various *interviewer observation variables*, recording characteristics about each household and surrounding neighbourhood. C*all record data* are also available for each wave. These data contain information about each visit to the household, including date and time of each call and the call outcome which is categorised into non-contact, contact, appointment, interview, and 'any other status' (this last category includes ineligibles and refusals and is defined in this way by the survey agency. This particular categorisation is not under our control). Call record data are defined for each household and are not available at the individual level, as is usual for most surveys. Further features of the survey and its sample design are discussed in detail in Durrant *et al*. (2015) and also in Buck and McFall (2012) and will not be repeated here.

2.2 Analysis sample and construction of the datafile

Since we are interested in nonresponse analysis in subsequent waves of a longitudinal study, the analysis sample conditions on response to wave 1 (wave 1 nonresponse -similar to nonresponse in a cross-sectional survey- was analysed in Durrant et al. (2015)). An advantage is that the survey variables from wave 1 (or alternatively from any previous wave) provide detailed information about both responding and nonresponding cases in wave 2 (or subsequent waves). The Understanding Society survey has, as a whole, multiple components: the General Population Sample (GPS), the Ethnic Minority Boost Sample (EMB), and the British Household Panel Study (BHPS) sample (McFall 2013). However, for this study only the main stage sample, the GPS, is of interest. (The BHPS sample did not take part in wave 1 and was mainly interviewed via telephone in subsequent waves. The EMB sample was excluded from the analysis as the rules for the selection of this sample are quite different from the main sample, and differences in sample selection for this subset are not of interest here.) Since we are interested in interviewer contact attempts, we focus on the face-to-face components in this study.

To construct the desired datafile at the household level the call record variables, interviewer variables and survey variables from wave 1 had to be linked to call record variables, interviewer observation variables and the final response outcome from wave 2. Understanding Society, like many other longitudinal surveys, does not include a unique household identifier that remains identical across waves for the same household (note that naturally for some households the household composition would be expected to change over time and it would not be feasible to allocate a stable household id number to all households). Hence, the unique identifier number at the individual level, which remains the same for the same individual across waves, had to be used for the linkage. In order to do this, first each member of each household in both waves

obtained the same sequence of calls as the whole household. Then, the two waves were merged on the basis of the unique individual identifier. Finally, the linked individuals were grouped back into households again (based on the household id number defined for wave 2 data), such that a household level analysis is possible. (It was possible to link all cases based on the individual identification number). The vast majority of households have the same household composition in both waves. Any households that had one (or more) individuals joining the household (from outside the survey) between waves 1 and 2 do not cause any concerns for the analysis (the models control for household composition and any indications of changes derived via the interviewer observation variables). There was a small number of households (159 or 1.5%) which split into two or three households between wave 1 and 2 and these were included in the analysis as two or three separate households in wave 2 (the models include a household split indicator). There was no case where two separate households interviewed in wave 1 formed one household in wave 2. The resulting (*initial) analysis sample* contained 24,896 households including households with sequences available in both waves and responding in wave 1.

The aim of the analysis is to predict the final response outcome in wave 2 given wave 1 survey and paradata information. In addition, we aim to investigate if the predictive power of the models can be improved if initial wave 2 call record data and interviewer observation data are available. The exploratory work conducted for this analysis suggested that information from the first three calls in the call sequences may be sufficient to reach an acceptable level of predictive power of the models (see also Durrant *et al*. 2015). Therefore (and to guarantee the comparability of the different models) the *final analysis sample* is restricted to all households from wave 2 that received more than three calls (11,029 households). This approach enables us to employ call record information from the first three calls in wave 2 to predict final length and

outcome of call sequences. The approach is relevant for survey practice as it helps to answer the question whether after a few number of call attempts (such as one, two or three calls) it is possible to predict the final outcome at a later call.

There are only a very small number (174) of missing cases in wave 1 in some of the geographic information and design variables since these are derived from administrative data. Date and time of a call are automatically captured using computer assisted methods leading to no missing cases in these variables. Recordings of the call outcome of the households of interest did not contain any missing information either. There was a small number of households with missing items in the wave 1 survey variables, and wave 1 and wave 2 interviewer observation variables and these cases were also deleted (399 cases or 3.6%). The *final analysis sample*, including only cases with four or more calls in wave 2, therefore, contains 10,630 households with information of interest from wave 1 and wave 2.

2.3 Response and explanatory variables

The key dependent variables are sequence length and response outcome in wave 2. We explored a range of different specification for the dependent variables such as different categorisation of sequence length (2, 3 or 4 category variables; requiring binary or multinomial logistic models) as well as defining sequence length as a count variable (requiring a Poisson regression model). The overall conclusions were very similar to the ones selected for presentation in this paper. The final analysis results, as presented here, are based on the following definitions of the three response variables:

1.) *length of call sequence* (binary), distinguishing short sequences (up to six calls) and long sequences (more than six calls). The cut-off point at six calls was intentionally selected to fit the survey protocol requirements of conducting a minimum of six call attempts if contact was not established earlier in the process.

2.) *(final) outcome of call sequence* (binary), distinguishing successful call sequences with at least one interview conducted in a household (after call 3) and unsuccessful call sequences with no interviews achieved (after call 3). (We recognise that a successful call sequence can be defined in several different ways, for example as all interviews achieved in a household. However, we chose this definition here since it is the least restrictive.)

3.) *a variable combining both length and final outcome,* distinguishing 4 categories - short unsuccessful (up to six calls, no interview in the sequence), short successful (up to six calls, at least one interview after call 3), long unsuccessful (more than six calls, no interview in the sequence) and long successful (more than six calls, at least one interview after call 3).

Table 1 presents the distributions of the three response variables used in the analysis.

[Table 1 about here]

The explanatory variables in the models can be split into seven main groups. (The distributions of the explanatory variables broken down by the categories of the three response variables used in the analysis are presented in the online Appendix Table A2. The exact wording of all variables and details of derived variables are provided in the online Appendix Table A1.)

1.) *geographic information and design variables from wave 1* (4 variables: urban/rural indicator, government office region, low density area for ethnic minorities, and month and year of household issue);

2.) *interviewer observation variables from wave 1* (12 variables, e.g. indicators of entry barriers, conditions of surrounding area such as litter in street, abandoned buildings, heavy traffic, type of accommodation, relative condition of the property, garden);

3.) *survey variables from wave 1* (many household level survey variables were explored and 9 variables were selected for the final analysis, e.g. household characteristics such as household income, the highest educational qualification, composition of the household);

4.) *call record variables* from wave 1 (6 variables, e.g. proportion of non-contact calls in a sequence or proportion of appointments in a sequence; length of sequence in wave 1; latent classes of sequences (see section 3 for their derivation));

5.) *call record variables* from wave 2 (11 variables, e.g. date, time of day, day of week, call outcome; also derived variables including time between calls);

6.) *interviewer observations from wave 2* (9 variables, e.g. household split identifier, presence of a car or a van, relative condition of the property, conditions of surrounding area such as litter in street, abandoned buildings, heavy traffic, type of accommodation, presence of children in a household, relative condition of the property, garden);

7.) *changes in interviewer observations between waves 1 and 2 identifiers* (8 derived variables indicating if there was a likely change between the observations between wave 1 and wave 2, e.g. change in conditions of the garden between the two waves, change in presence of a car or a van between the two waves).

3. Analysis approach

First, the main analysis strategy is described. Then, the response outcome variables and resulting model specifications are defined more formally. Different model specifications are considered, including latent class analysis. A range of assessment and evaluation criteria are presented to guide comparisons between models.

3.1 Response variables, model specifications and modelling strategy

The response variables were introduced in the data section and include length of call sequence (short versus long sequence) and response outcome (successful versus unsuccessful sequence). These distinctions are motivated by research questions relevant to survey practitioners. To save costs, survey practice is interested in identifying cases early on (i.e. solely based on previous wave information or after just a few calls) which are likely to have an unsuccessful response outcome and which take a long time to respond. We are therefore interested in identifying households (or groups of households) that are likely to have *long and unsuccessful* call sequences. More formally, we employ the following three dependent variables and resulting binary logistic and multinomial models.

We denote by y_i the (binary or multinomial) response variable of household i. The dependent variable *length of call sequence* is defined as

$$y_i = \begin{cases} 1 & \text{short call sequence (up to 6 calls)} \\ 0 & \text{long call sequence (more than 6 calls)} \end{cases},$$

and the *final outcome of call sequence* is coded

$$y_i = \begin{cases} 1 & \text{successful call sequence (at least one interview)} \\ 0 & \text{unsuccessful call sequence (no interview).} \end{cases}$$

For the two dependent variables, the response probabilities are denoted by $\pi_i = Pr(y_i = 1)$ and are related to the explanatory variables using logistic regression (e.g. Agresti 2013):

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad (1)$$

where $\boldsymbol{x}_i$ is a vector of household-level covariates including intercept and interactions, and $\boldsymbol{\beta}$ is a vector of coefficients.

Combining both length and outcome we define the third dependent variable as

$$y_i = \begin{cases} 1 & \text{short successful (up to 6 calls, at least one interview)} \\ 2 & \text{short unsuccessful (up to 6 calls, no interview)} \\ 3 & \text{long successful (more than 6 calls, at least one interview)} \\ 4 & \text{long unsuccessful (more than 6 calls, no interview).} \end{cases}$$

For this dependent variable multinomial logistic regression is used. If the response variable has S categories, then the multinomial logistic regression model can be expressed as a set of S-1 non-redundant logistic model equations. The response probabilities are denoted by $\pi_i^{(s)} = Pr(y_i = s), s = 1, 2, 3, 4$. Taking 'long unsuccessful' as the reference category, the multinomial logistic regression model can be expressed as

$$log\left(\frac{\pi_i^{(s)}}{\pi_i^{(4)}}\right) = \boldsymbol{\beta}^{(s)^T} \boldsymbol{x}_i^{(s)}, s = 1, 2, 3, \qquad (2)$$

where $\boldsymbol{x}_i^{(s)}$ is a vector of covariates including intercept and interactions, and $\boldsymbol{\beta}^{(s)}$ is a vector of coefficients.

To allow for comparison of predictability of all models and based on our analysis sample all call outcomes are with reference to after the first 3 calls (the measures of goodness-of-fit and predictability of the models as outlined in section 3.3 allow the comparison of different models

for the same data only). (When applying the analyses methods in practice to find the best predictive model for any type of analysis sample, this restriction is not necessary. Here, it is only used to allow strict comparisons between models.) The cut-off point of six calls reflects the protocol of the data collection process which suggests that each household should have at least six calls if productive calls were not obtained earlier in the process (McFall 2012). As already mentioned in the data section (section 2.3) different specifications of the dependent variables (including different cut-off points, different number of categories such as 2, 3, or 4 categories, also definition as a count variable, requiring a range of binary and multinomial logistic and Poisson models) were explored with no significant changes in the key findings. In all models robust standard error estimation is used to correctly account for the clustering of households within interviewers (Huber 1967; White 1980, 1984 and 1994). The models allow for the primary stratification present in the survey by including geographical stratification variables into all models. Likelihood ratio tests (using the change in the L^2 goodness-of-fit statistic) are used to test the significance of a term in a model. A forward stepwise model selection procedure was employed. Explanatory variables are included into the models by groups discussed earlier: first, only geographic and design variables from wave 1 are included; then survey variables from wave 1 are added, followed by interviewer observations variables from wave 1. At the next step of the model building procedure, call record data from wave 1 are added. Then, interviewer observations from wave 2 and indicators of change in interviewer observations between the two waves are introduced. At the last stage of modelling, call record data from the current wave, including call outcomes from the first three calls in wave 2 are added to the final models.

3.2 Latent class analysis

In order to control for call histories in previous waves, different summary measures can be produced. One approach, that so far has not yet been used in the context of response prediction, is latent class analysis (LCA). The resulting summary measure is then used as an explanatory variable in the models. LCA is a model-based technique which allows summarising data in the form of one latent variable without significant loss of information (Bartholomew *et al*. 2008; Hagenaars and McCutcheon 2002). LCA helps to split a heterogeneous sample into classes which are more homogenous. The main aim of the LCA is to determine the smallest number of classes that is sufficient to explain relationships between manifest variables (Magidson and Vermunt 2004).

For example, if there are six manifest or observed variables (A, B, C, D, E and F), then the latent class model can be expressed as

$$\pi_{ijklmnt} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X} \pi_{mt}^{E|X} \pi_{nt}^{F|X}, \qquad (3)$$

where $\pi_{ijklmnt}$ is the probability that response *i* is obtained for item *A*, response *j* for item *B*, response *k* for item *C*, response *l* for item *D*, response *m* for item *E*, response *n* for item *F* and is in latent class *t* of a latent variable *X*; π_t^X denotes the probability of being in the latent class $t = 1,2,\ldots,T$ of the latent variable *X*; $\pi_{it}^{A|X}$ denotes the conditional probability of obtaining response to item *A*, from members of class *t*, $i=1,2,\ldots,I$; and $\pi_{jt}^{B|X}$, $\pi_{kt}^{C|X}$, $\pi_{lt}^{D|X}$, $\pi_{mt}^{E|X}$, $\pi_{nt}^{F|X}$ with $j=1,2,\ldots,J$, $k=1,2,\ldots,K$, $l=1,2,\ldots L$, $m=1,2,\ldots,M$, $n=1,2,\ldots,N$ denote the corresponding conditional probabilities for items *B*, *C*, *D*, *E* and *F* respectively (Magidson and Vermunt, 2004). In our

analysis, the manifest variables are the call outcomes for the first six calls in wave 1. (Different numbers of calls were also explored but the overall conclusions were very similar).

In order to determine the number of homogeneous classes, which exists in the heterogeneous population with respect to the latent variable, the model fit should be assessed. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) provide reliable measures of the model fit and help to determine the number of classes (Akaike 1974; Bozdogan 1987; Magidson and Vermunt 2004; Muthén 1998-2004; Schwarz 1978), and both measures are used in this analysis. The AIC is the goodness-of-fit statistic corrected for the complexity of the model by taking into account the number of parameters which were estimated (Field 2009). The BIC is similar but more conservative than the AIC (Field 2009). This statistic balances two components of a model, the likelihood value and parsimony (Muthén and Muthén 2000). For both criteria smaller values represent a better fit of the model (Dias 2001; Field 2009).

The level of potential classification error or classification quality is also important to consider when deciding on the final model (Muthén and Muthén 2012; Storr *et al*. 2004). According to Beadnell *et al*. (2003), classification quality is the ability to distinguish membership in the latent class given the model and the data. The higher the average class probabilities the better the ability to accurately classify sequences into their classes (Beadnell *et al*. 2003). According to Storr *et al*. (2004), model fit can be improved by adding more latent classes, but this additional class may make the model less interpretable. Therefore, it is important to use the judgment and the principal of parsimony when deciding on the final model. Once the decision about the number of classes is taken, sequences are allocated to the appropriate latent classes on the basis of the call outcomes in the six calls with the help of estimated posterior probabilities. The posterior probability is the probability of a sequence being in the latent class *t*

given a specific response pattern in a particular sequence (Bartholomew *et al*. 2008). Once the posterior probabilities are estimated, every sequence of calls in the dataset can be assigned to a particular class for which the posterior probability is the highest (Magidson and Vermunt 2004). This new variable which contains classes to which sequences were allocated is then used as an explanatory variable in the analysis. In order to conduct latent class analysis and to obtain the latent classes for the sequences, Mplus 7 statistical software package is employed (Muthén and Muthén 2012).

3.3 Comparison of model performance and evaluation

The standard way of assessing the model performance of nonresponse models in the literature is to use the (pseudo) R^2 statistic (Groves and Couper 1996; Bates *et al*. 2008; Olson and Groves 2012; Olson *et al*. 2012; West and Groves 2013), which is a goodness-of-fit statistic representing the proportion of variation in the dependent variable that is explained by the model. The closer the statistic is to 1, the greater the proportion of variation explained by the model. However, as pointed out in Plewis *et al*. (2012) this is not the most appropriate measure to evaluate the ability of a model to predict the outcome. In particular, it does not distinguish between the accuracy of the model for nonrespondents and respondents. Instead, several measures are proposed, that are used for model comparisons (see also Altman 1991; Pepe 2003; Plewis *et al*. 2012; Agresti 2013; Durrant *et al*. 2015): discrimination and prediction, classification table values (the proportion of correctly classified cases), measures of sensitivity and positive predicted values, and the area under the curve (AUC) of the ROC.

Let us start with the binary case. Let $\hat{y}_i$ denote the predicted value for an observation *i*, and $\hat{\pi}_i$ the predicted response propensity from the model. The predicted value is obtained

depending on a cut-off π_0, i.e. the prediction for observation *i* is $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi_0$, and $\hat{y}_i = 0$ if $\hat{\pi}_i \leq \pi_0$. (The default options for setting π_0, which are also used here initially, are $\pi_0 = 0.50$ for the binary and $\pi_0 = 0.25$ for the multinomial case, although in practice different values can be explored. We also allow for all possible values by using ROC curves, see below). *Classification tables* are obtained by cross-classifying the observed binary response, y_i , with the predicted values, $\hat{y}_i$, i.e. classification tables allow the evaluation of the two concepts: discrimination and prediction. *Discrimination* is simply the conditional probability that a case is predicted to be a respondent (nonrespondent) given that a household is indeed a respondent (nonrespondent). Formally, discrimination can be expressed as $P(\hat{y}_i = 1|y_i = 1)$ (referred to as *sensitivity*) and $P(\hat{y}_i = 0|y_i = 0)$ (*specificity*). *Prediction* describes the conditional probability of being a respondent (nonrespondent) given a household is predicted to be a respondent (nonrespondent), which can be expressed formally as $P(y_i = 1|\hat{y}_i = 1)$ (*positive predictive value)* and $P(y_i = 0|\hat{y}_i = 0)$ (*negative predictive value*). The concept of prediction is particularly useful for our research questions here, since the true outcomes are not actually observed until data collection has been completed and hence survey researchers are interested in the ability of a response model to predict the true outcome correctly, given the predicted values from the model. Another useful measure, that can be derived from the classification table, is the percentage of observations correctly classified, which is an overall summary measure of model performance, and reflects the summary of the diagonal of the classification table as a weighted average of sensitivity and specificity:

$$P(\text{correctly classified}) = P(y_i = 1 \text{ and } \hat{y}_i = 1) + P(y_i = 0 \text{ and } \hat{y}_i = 0)$$

$$= P(\hat{y}_i = 1|y_i = 1)\, P(y_i = 1) + P(\hat{y}_i = 0|y_i = 0)\, P(y_i = 0)$$

In the results section we refer to sensitivity and positive predicted values with respect to modelling long and unsuccessful call sequences.

For the multinomial case, classification tables and therefore discrimination and prediction, can be similarly defined. Here we have several categories of correctly classified and misclassified cases. For a 4 category variable, as is the case in this paper, this results in a 4×4 classification table, allowing for 4 correctly (the diagonal) and 12 incorrectly classified groups.

A potential restriction is the dependency of prediction and discrimination (and therefore of classification tables) on the (arbitrary) cut-off value π_0. ROC curves (Agresti 2013) address this problem by deriving different measures across all possible cut-off values. The ROC curve plots sensitivity as a function of (1-specificity) for all possible π_0. For a given specificity, better predictive power corresponds to higher sensitivity. If π_0 is near 0, then most predictions are 1, which implies that sensitivity is near 1, specificity is near 0, and the point (1-specificity; sensitivity) is close to (1;1). If π_0 is near 1, almost all predictions are 0, then, sensitivity is near 0, specificity is near 1, and (1-specificity; sensitivity) is close to (0;0). To help interpretation, the higher the ROC curve, i.e. the greater the AUC, the better is the predictive power of the model.

4. Results

Table 2 presents a range of models starting from the basic model, only controlling for geographic and design variables, up to a model that controls for previous and current wave paradata, interviewer observations, survey variables and the outcome of the most recent calls. All modelling steps are carried out for the two binary logistic models (sequence length, final response outcome) and the joint multinomial model (sequence length and final response outcome). More than 25 models were fitted for each of the three dependent variables, exploring a

variety of model specifications, including different explanatory variables. Table 2 presents 8 selected models for each of the three different response outcomes. A range of assessment criteria are presented, including the pseudo R^2 value (Nagelkerke R^2 statistic (Nagelkerke 1991)), the percentage of the overall correctly classified values (derived from the diagonal of the classification table) and the AUC from the ROC curve. We are interested in models with a high pseudo R^2 value. The closer the pseudo R^2 value is to 1 the better is the goodness-of-fit of the model and the higher the proportion of variability in the response variable that is explained by the model. For comparison, standard response propensity models reported in the literature often have pseudo R^2 values of between 3-8% (Olson *et al*. 2012; Olson and Groves 2012; West and Groves 2013). When comparing models based on the values from the classification table, we are interested in those with higher values, indicating that a higher percentage of cases is correctly classified. (These values do not yet distinguish between the predictions of different categories. These results are presented in Tables 3-5.) To be able to interpret the classification table values in a meaningful way we compare them with the observed outcome distributions from wave 2 provided in Table 1. For comparison, without any prior information for the two binary outcomes (response and sequence length) we would expect about 50% of cases to be predicted correctly. For the multinomial outcome with 4 categories it would be 25%. With prior information, for example based on the observed outcome distributions from wave 2 provided in Table 1, comparing the values with the most frequently observed distribution in this table, we would expect about 63% for the variable length to be correctly classified, 71% for the variable outcome and 50% for the combined outcome of both length and response. We therefore aim to find classification values of above 63%, 71% and 50% respectively. The larger the differences between these base values and the values obtained for the models, the higher the predictive

power of the model. With respect to the AUC values we are interested in models with values above 0.5, indicating that a model classifies the group better than chance. (As an example, the ROC curves for the final model (Model 8) for length and for response outcome are given in the Appendix (Figures A1 and A2).)

The results for Model 1, the 'base model', indicate a very low pseudo R^2 value (2%) (Table 2). Although the classification table values are doing better than chance, there is no improvement when compared to the marginal distribution (63%, 70.7% and about 50%). The AUC values are low of just above 0.5. This suggests that geographic and design information on their own do not help in predicting the variables of interest in comparison to chance, despite the significance of some variables in the three models (month of survey, type of residence (urban/rural)). The standard approach in the nonresponse literature is to condition on the survey variables from the previous wave (or, if available, on any other fully observed variables such as from Census, register or administrative data). We therefore refer to Model 2, which includes survey variables from the previous wave, as the 'standard model'. Interestingly, survey variables improve the predictability of the models only very slightly (pseudo R^2 values now between 5-8%, classification table values are around 63%, 71% and 50% and the AUC is 0.62) despite many variables being highly significant. We find a very similar trend for Model 3, which conditions on interviewer observation variables from wave 1, indicating that although some of the interviewer observation variables are highly significant, they do not improve the actual prediction substantially. Models with historical call record information, including summary measures of call record data from the previous wave (e.g. proportion of noncontacts etc., with or without length of sequence which was added as a categorical and as a continuous variable), improve the models further but again this improvement is not very large (Model 4) (pseudo R^2 value is now

between 7% and 11%, the classification table values are 65%, 71% and 50%, the AUC is 0.64). Another way of accounting for historical call record data is to perform LCA. The best solution obtained in LCA using AIC, BIC and classification quality criteria contained 4 classes, with the classes as follows: the first class of sequences had 1-2 calls only, the second had sequences with 3 calls, the third had sequences with 4-5 calls and the fourth class had a high proportion of noncontact calls in the first 6 calls. Comparing the LCA approach (Model 4b) with the approach of using summary measures of historical call record information (Model 4) we can see that both approaches produce very similar results. The LCA approach, although worth exploring, does not seem to perform any better than controlling for the simple summary measures of historical call record data such as proportion of noncontacts or proportion of contacts for our data. The next step is to include paradata from the early stages of the current wave (wave 2), comprising interviewer observation variables (Model 5; this model also includes an indicator if a household split between the two waves) and an indicator if there has been a likely change between interviewer observations between the two waves (Model 6). Again the model performance is improved (the pseudo R^2 value is now 10%, 14% and 17%, the classification table is 65%, 73% and 52%, and all AUC are around 0.7). Including also call record variables (such as timings of calls and time between calls) again leads to an improvement with pseudo R^2 values now reaching 11%, 18% and 22%, and the classification table reaching 66%, 74% and 53% for the first time). The best analysis results are achieved for the final model (Model 8), which includes the outcome of the last 3 calls in the current wave (the pseudo R^2 value reaches 24%, 27% and even 36%, classification table values of 70%, 77% and 56% and all AUC above 0.75, which is significantly larger than 0.5). The values are now clearly higher than for standard nonresponse models (see also Plewis *et al*. 2012), meaning that discrimination between respondents and nonrespondents is

better. From Table 2, we can see that the final model (Model 8) is significantly better than the base model (Model 1) and the standard nonresponse model (Model 2) for all three types of models (binary and multinomial). Exploring the final model further, controlling for the outcomes one at a time, we find the more recent the call outcome information, the better is the performance. The outcome of the most recent call contributes to the biggest improvement in comparison to, for example, including the outcome of just the first call or the first two calls (results not shown).

The values of the classification table given in Table 2 provide information on the overall probability of correctly classified cases. However, this overall measure does not provide an indication of how well we are classifying the values with respect to particular groups, such as the long unsuccessful call sequences, which is the group of our primary interest. To start with, Table 3 indicates for all 8 models the results of the discrimination power for the two binary and the multinomial modelling case, i.e. the percentage of correctly classified households by categories of the dependent variables (i.e. for the two binary cases sensitivity $P(\hat{y}_i = 1|y_i = 1)$ and specificity $P(\hat{y}_i = 0|y_i = 0)$ are shown and for the multinomial case $P(\hat{y}_i = s|y_i = s)$, for $s = 1,2,3,4$). The results clearly show that the base model and the standard model are not performing very well, since, in fact, they predict (almost) all outcomes as short successful and do not discriminate between the different categories. Although, as we have seen in Table 2, this leads to a relatively high percentage of overall correctly classified cases, the models perform in reality very poorly with regards to our category of interest, the long unsuccessful cases. We can see that, broadly speaking, the more sophisticated the models become, the better their performance. For example, including prior wave call record data and interviewer observation variables, increases the discrimination power to about 20% (long), 7% (unsuccessful) and 10% (long unsuccessful). For models including paradata from the current wave (Models 5-8), this

increases to just above 20% for the multinomial case (Models 5-7) and even to 31% for the final multinomial model, including the outcomes of the most recent calls (Model 8), meaning that about 31% of households that have long unsuccessful call sequences 31% are correctly classified by the model as being indeed in this category. (For the two binary cases these are 50% for the long and 37% for the unsuccessful categories respectively.)

[*Table 3 about here*]

From a survey practice perspective, another, possibly even more useful, measure is the ability to predict the different outcomes well (rather than to discriminate between the different categories). This means in practice, that if the model predicts a particular outcome for a household, such as a long unsuccessful call sequence - either before wave data collection starts or after just one, two or three calls - the prediction measure gives us the probability of indeed identifying a true long unsuccessful outcome. Table 4 shows the predictive power for all three types of models and for all 8 modelling stages (i.e. for the two binary dependent variables the table shows $P(y_i = 1|\hat{y}_i = 1)$ (positive predictive value) and $P(y_i = 0|\hat{y}_i = 0)$ (negative predictive value) and for the multinomial case ($P(y_i = s|\hat{y}_i = s)$), for $s = 1,2,3,4$). We can see clearly from Table 4 that the base model again performs very poorly, not predicting any cases correctly as long unsuccessful calls. Interestingly, the standard nonresponse model is already a good improvement predicting about 53% of the long, 56% of the unsuccessful and 35% of the long unsuccessful cases correctly. The values improve slightly for models including historic paradata. They improve further when most recent paradata are included. The final model (Model 8) again indicates the best performance with above 60% (long), 65% (unsuccessful) and 41% (long unsuccessful). Summarising all results from Tables 2, 3 and 4 we conclude that the models including historic paradata improve the prediction of the base and standard model. The most

recent paradata (paradata from the current wave), in particular the outcome of the most recent calls, are the most useful predictor variables in the models.

[*Table 4 about here*]

Given that it is not possible to predict all cases correctly, in a final step we are interested in how the cases, that the model did not predict correctly, are distributed. Table 5 breaks down further the modelling results for Model 8 of the multinomial model from Tables 3 and 4, now showing the complete classification table. The upper panel (panel A) indicates sensitivity and the lower panel (panel B) shows the positive predicted values. (Note that the diagonals in Table 5 for cases A and B are the last row from Table 3 and 4 for the multinomial model respectively.) We are particularly interested in panel B, the case where the model predicts a long unsuccessful outcome. We can see that 41.4% are indeed in this category (the same result was already reported in Table 4), and for the remaining cases 20.1% and 19.1% are classified as short successful and short unsuccessful respectively. It should be noted that the misclassification to short sequences (successful or unsuccessful) would not have in practice negative implications since the recommended 6 calls might be made anyway. We can see that actually only 19.3% are classified incorrectly as long successful.

[Table 5 about here]

The analysis identified a range of variables as significant or highly significant across the various models of interest. Although we do not wish to go into detail with the discussion of the coefficients in the models some of the main findings are briefly highlighted. The full modelling results for the final models (Model 8) for the two binary outcomes and the multinomial outcome are given in the online Appendix (Tables A3 and A4). Although the inclusion of survey variables

as seen earlier does not improve the ability of the models to predict the outcomes of interest by very much, many of the variables are significant or highly significant (e.g. highest qualification), supporting well known correlates of nonresponse in longitudinal surveys (Lepkowski and Couper 2002; Watson and Wooden 2009; de Leeuw and de Heer 2002; Campanelli and O'Muircheartaigh 1999; Pickery *et al*. 2001; Haunberger 2010). We noted earlier that paradata from the previous wave increase the ability of the models for prediction. Indeed, we find a range of interviewer observation variables (including derived indicators of changes in households between waves) and call record variables to be (highly) significant across the range of models. Interestingly, we observe that households that had long sequences in the previous wave, a high proportion of noncontact calls or a high proportion of calls with contacts (but no further outcomes) are indeed significantly more likely to also have long call sequences in the current wave. This indicates that trends over time (across waves) may indeed exist and some households with a particular calling pattern may exhibit a similar calling pattern in a future wave. Whilst some variables are significant for both length and response outcomes, others only predict one of the dependent variables (e.g. the variable length of sequence in previous wave does not have a significant impact on call outcome, whereas it does predict length of call sequence; also, if the household has people of pension age then this has a highly positive impact on sequence length (predicting a short call sequence), whereas it is not significant in the outcome variable; times between calls are significant for both length and response outcome). In addition, some variables are found to be significant across all of the various modelling stages (across models 1-8) indicating consistent influences on the dependent variables, whereas others are sometimes significant and sometimes not (for example time between calls was significant across all of the

modelling stages for final response outcome, whereas time of day was sometimes significant and sometimes not, indicating an unstable relationship with the dependent variables).

5. Conclusions and implications for survey practice

This paper aims to use paradata and survey data from the previous and current wave of a longitudinal study to improve the prediction of nonresponse models and to use the resulting models for informing current and future survey designs. Although the use of paradata in nonresponse modelling has increased in recent years (Potthoff *et al*. 1993; Groves and Couper 1996; Bates *et al*. 2008; Kreuter *et al*. 2010a and 2010b; Sinibaldi *et al*. 2013; Sinibaldi *et al*. 2014; Wagner 2013a and 2013b; Durrant *et al*. 2015), it is yet unanswered if historic and/or current paradata are useful in the context of a longitudinal survey. Whilst so far nonresponse modelling has focussed on the final response outcome or on outcome at the next call (Groves and Heeringa 2006; Durrant *et al*. 2011; Durrant *et al*. 2013a and 20134b; Hanly 2014; Sinibaldi 2014; Durrant *et al*. 2015), the models presented predict both sequence length and response outcome, separately and jointly. The prediction of particular types of call outcomes, such as long and unsuccessful call sequences, are assessed. This approach may be particularly useful from a survey practice perspective: if we are able to predict, for example, long unsuccessful call outcomes before data collection or after just a few calls (such as one, two or three calls) it may be possible for survey practitioners to make informed decisions about future tailored treatment approaches, either by stop calling or by allocating increased data collection efforts to obtain a response from more difficult households. Standard nonresponse models are often developed for understanding influences on nonresponse better (i.e. analysis of the significance of correlates in

the model is of interest, see for example Durrant and Steele 2009) or for nonresponse adjustment, such as for the development of a weighting model. To be able to predict response outcomes with the aim of changing current or future survey designs different assessment criteria need to be used. In addition to the standard approach of the (pseudo) R^2 statistic, this paper proposes the use of classification tables, discrimination (sensitivity and specificity), prediction (positive and negative predicted value) and the AUC of the ROC curve. The paper also explores different model specifications and the inclusion of a range of specification of explanatory variables, including variables derived via a latent class modelling approach.

In the following, the most important findings for both survey methodology and survey practice are summarised:

1. The findings indicate that 'basic' models (including geographic and design variables) and 'standard' nonresponse modelling approaches (only accounting for previous wave survey data) although commonly used in predicting and adjusting for nonresponse, do not predict the response outcome very well (R^2 values are between 5-8% which is to be expected for standard nonresponse models (Olson *et al*. 2012; Olson and Groves 2012; West and Groves 2013), the classification table values of the percentage of correctly classified cases are 63%,71% and 50% depending on the type of model, better than chance but not better than the observed distribution).
2. Conditioning on previous wave paradata, including call record data, interviewer observation data and indicators of change, improve the fit of the model. A significant improvement can be observed when conditioning on current wave paradata (from the initial stages of the current wave data collection), in particular when conditioning on the most recent call outcome (pseudo R^2 values reach 24%, which is very high in a social

science context; of the long call sequences we can predict approximately 60% correctly, of the unsuccessful call sequences 65% and of the long and unsuccessful cases 41% correctly, which is much higher than for the standard models and estimation by chance). The findings may indicate that the nonresponse process predominantly depends on the most current circumstances of a sample unit and may be less determined by past events. The fact that overall it is difficult to predict nonresponse, may also indicate that the nonresponse process in parts may be a random process, which is difficult to predict by nature.

3. A latent class analysis approach provides an attractive way of taking account of historic call record data into the models. For our data, we find that the latent class analysis approach performs very similarly to an approach of including derived simple summary measures into the models. Also, different model specifications (e.g. different specifications of the dependent variables) did not alter the main conclusions of the findings.
4. Several interviewer observation variables (including derived indicators of changes in households between waves) and call record variables are found to be (highly) significant across the range of models. Interestingly, we observe that households that had long sequences in the previous wave, a high proportion of noncontact calls or a high proportion of calls with contacts (but no further outcomes) are indeed significantly more likely to also have long call sequences in the current wave. This indicates that trends over time (across waves) may indeed exist and some households with a particular calling pattern may exhibit a similar calling pattern in a future wave.

5. The results for the different assessment criteria of the models have shown that it is worthwhile exploring a range of methods to evaluate and compare the models. The commonly used approach of the R^2 statistic alone is not sufficient in this context. Concepts frequently used in epidemiology, such as discrimination, prediction and AUC are recommended (see also Plewis *et al*. 2012). These allow the assessment of the ability of the models to predict certain groups of the dependent variables, which is of interest here, such as predicting long unsuccessful calls.
6. Often, significance of variables in a model is used as an indication that controlling for such variables improves the fit of the model. Many variables have been found to be highly significant in the models considered and these include a range of interviewer observation variables, call record variables (previous and current) and survey variables. However, prediction can still be low depending on the model. Therefore, significance of correlates in a model alone is not sufficient to assess the predictive power of the model and its use for adaptive and responsive survey designs.
7. In this paper we also find that modelling call sequence length in addition to just the response outcome which is common in the nonresponse literature helps in understanding future calling patterns.

Currently the work does not take account of any cost data (and these data are also not available to us). In practice some calls may be relatively inexpensive, whereas other types of calls or visits may be more burdensome for the survey agency. For example, a call to a household on the way to another household may be carried out at relatively little cost. Survey researchers may wish to take this type of information into account when making decisions on which households best to follow up or when to stop calling. It should be noted that the study here

uses observed interviewer calls. The data were not obtained by a random allocation of interviewers or interviewer calls to households. Hence, it is possible to analyse associations but causal statements cannot be made. However, this is not a limitation since we are interested in the comparison of different models and in identifying indicators that help to predict future outcomes as it would be the case in a standard survey design setting (rather than in an experiment). The data used here do not include any feedback variables from interviewers (for example interviewer ratings on how likely the case is to respond and when). We are aware of only one other study in this area (Eckman *et al.* 2013). In future work, it would be of interest to assess the ability of the models to predict the outcomes when such interviewer assessment variables are included, using the evaluation criteria of discrimination and prediction outlined in this paper.

It is hoped that the modelling and assessment approach presented here will help survey practitioners to improve nonresponse models and prediction to inform current and future survey design decisions. As was already pointed out in Plewis et al. (2012) we strongly recommend the use of discrimination, prediction, classification tables and ROC curves rather than simply the (pseudo) R^2 value to assess predictability of response models. The methodology outlined in this paper can be used and adapted by survey managers of other datasets. The approach is currently implemented by Statistics Sweden in an adapted form to the Swedish Labour Force Survey, to help cut costs of unproductive interviewer telephone calls and personal visits to households.

6. References

Agresti, A. 2013. *Categorical data analysis.* New Jersey: John Wiley & Sons.

Altman, D. G. 1991. *Practical statistics for medical research.* London: Champan & Hall.

Akaike, H. 1974. “A new look at statistical model identification.” *IEEE Transactions on Automatic Control* AC-19: 716-723. DOI: 10.1109/TAC.1974.1100705.

Bartholomew, D., F. Steele, I. Moustaki, and J. Galbraith. 2008. *Analysis of multivariate social science data*. London: CPC Press.

Bates, N., J. Dahlhamer, and E. Singer. 2008. “Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse.” *Journal of Official Statistics* 24: 591-612.

Beadnell, B., S. Baker, K. Knox, S. Stielstra, D.M. Morrison, E. DeGooyer, L. Wickizer, A. Doyle, and M. Oxford. 2003. “The influence of psychosocial difficulties on women’s attrition in an HIV/STD prevention program.” *AIDS Care* 15(6): 807-820. DOI**:** 10.1080/09540120310001618658.

Biemer, P. P., P. Chen, and K. Wang. 2013. “Using level-of-effort paradata in non-response adjustments with application to field surveys.” *Journal of Royal Statistical Society: Serie* A 176: 147–168. DOI: 10.2307/23355181.

Bozdogan, H. 1987. “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions.” *Psychometrika* 52: 345-370. DOI: 10.1007/BF02294361.

Buck, N. and S. McFall. 2012. “Understanding society: design overview.” *Longitudinal and Life Course Studies* 3 (1): 5-17.

Campanelli, P. and C. O’Muircheartaigh. 1999. “Interviewers, interviewer continuity, and panel survey nonresponse.” Quality and Quantity 33: 59–76. DOI: 10.1023/A:1004357711258.

De Leeuw, E., and W. de Heer. 2002. “Trends in household survey nonresponse: A longitudinal and international comparison.” In *Survey nonresponse* edited by R.M. Groves, D.A. Dillman, J.L.Eltinge, and R.J.A. Little, 41-54. New York: John Wiley and Sons.

Dias, J.G. 2001. *Components of knowledge on AIDS in Brazil: Identifying information needs using a segmented approach*. Population Research Centre Working Paper. University of Groningen: Population Research Centre.

Durrant, G. B., J. D’Arrigo, and G. Müller. 2013a. “Modeling call record data: Examples from cross-sectional and longitudinal surveys.” In *Improving surveys with paradata: Analytic uses of process information,* edited by F. Kreuter, 281-308. New Jersey: Wiley and Sons.

Durrant, G. B., J. D'Arrigo, and F. Steele. 2011. "Using field process data to predict best times of contact conditioning on household and interviewer influences." *Journal of Royal Statistical Society: Series* A 174: 1029-1049. DOI: 10.1111/j.1467-985X.2011.00715.x.

Durrant, G. B., J. D'Arrigo, and F. Steele. 2013b. "Analysing interviewer call record data by using a multilevel discrete-time event history modelling approach." *Journal of Royal Statistical Society: Series* A 176: 251-269. DOI: 10.1111/j.1467-985X.2012.01073.x.

Durrant, G. B., O. Maslovskaya, and P. W. F. Smith. 2015. "Modelling final outcome and length of call sequence to improve efficiency in interviewer call scheduling." *Journal of Survey Statistics and Methodology* 3: 397-424. DOI: 10.1093/jssam/smv008.

Durrant, G.B., O. Maslovskaya, and P.W.F. Smith. 2016. "Investigating call record data using sequence analysis to inform adaptive survey designs**.**" National Centre for Research Methods (NCRM) Working Paper. University of Southampton, (submitted).

Durrant, G.B. and F. Steele. 2009. "Multilevel modelling of refusal and noncontact nonresponse in household surveys: Evidence from six UK government surveys." *Journal of the Royal Statistical Society: Series A* 172(2): 361-381. DOI: 10.1111/j.1467-985X.2008.00565.x.

Eckman, S., J. Sinibaldi, and A. Montmann-Hertz. 2013. "Can interviewers effectively rate the likelihood of cases to cooperate?" *Public Opinion Quarterly* 77(2): 561-573. DOI: 10.1093/poq/nft012.

Field, A. 2009. *Discovering statistics using SPSS*. Los Angeles: SAGE.

Groves, R. M., and M.P. Couper. 1996 "Contact-level influences on cooperation in face-to-face surveys." *Journal of Official Statistics* 12: 63–83.

Groves, R. M. and S. G. Heeringa. 2006. "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of Royal Statistical Society: Series* A 169: 439-459. DOI: 10.1111/j.1467-985X.2006.00423.x.

Hagenaars, J.A., and A.L. McCutcheon. 2002. *Applied latent class analysis*. Cambridge: Cambridge University Press.

Hanly, M. 2014. "Improving nonresponse bias adjustments with call record data." *Conference paper*. 25th International Workshop on Household Survey Nonresponse, Iceland.

Hanly, M., P. Clarke and F. Steele 2015. "Sequence analysis of call record data: Exploring the role of different cost settings". *Journal of the Royal Statistical Society: Series A (Statistics in Society).* DOI: 10.1111/rssa.12143.

Haunberger, S. 2010. “The effects of interviewer, respondent and area characteristics on cooperation in panel surveys: a multilevel approach.” *Quality and Quantity* 44: 957–969. DOI: 10.1007/s11135-009-9248-5.

Huber, P. J. 1967. “The behavior of maximum likelihood estimates under nonstandard conditions.” In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA: University of California Press, vol. 1: 221–233. Available at: https://projecteuclid.org/euclid.bsmsp/1200512988 (accessed 31 March 2016).

Kreuter, F. (ed.) 2013. *Improving surveys with paradata: Analytic uses of process information*. New Jersey: Wiley and Sons.

Kreuter, F., M. Couper, and L. Lyberg. 2010a. “The use of paradata to monitor and manage survey data collection.” In *Proc. of the Joint Statistical Meeting, Section of Survey Research Methods*, 282-296. Vancouver, Canada.Kreuter, F. and U. Kohler. 2009. “Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey.” *Journal of Official Statistics* 25: 203-226.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R.M. Groves, and T.E. Raghunathan. 2010b. “Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys.” *Journal of Royal Statistical Society: Series* A 173: 389–407. DOI: 10.1111/j.1467-985X.2009.00621.x.

Lagorio, C. 2015. “Call and response: Modelling longitudinal contact and cooperation using lagged contact records data.” Paper presented at the conference of the European Survey Research Association (ESRA), Iceland, July 2015.

Lepkowski, J.M. and M.P. Couper. 2002. “Nonresponse in the second wave of longitudinal household surveys.” In *Survey nonresponse* edited by R.M. Groves, D.A. Dillman, J.L.Eltinge, and R.J.A. Little, 259-272. New York: John Wiley and Sons.

Magidson, J., and J.K. Vermunt. 2004. “Latent class models.” In *The SAGE handbook of quantitative methodology for social sciences*, edited by D. Kaplan, 175-198. Thousand Oaks: SAGE publications.

McFall, S. L. (ed.) 2012. *Understanding Society: Findings 2012*. Colchester: Institute for Social and Economic Research, University of Essex.

McFall, S. L. (ed.) 2013. *Understanding Society –UK Household Longitudinal Study: Wave 1-3, 2009-2012, User Manual*. Colchester: University of Essex.

Muthén, B.O. 1998-2004. *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén. Available at: http://www.statmodel.com/download/techappen.pdf (accessed 10 February 2016).

Muthén, B.O., and L.K. Muthén. 2000. "Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes." *Alcoholism: Clinical and Experimental Research* 24(6): 882-891. DOI: 10.1111/j.1530-0277.2000.tb02070.x.

Muthén, L.K., and B.O. Muthén. 2012. *Mplus user's guide*. Los Angeles, CA: Muthén and Muthén.

Nagelkerke, N.J.D. 1991. "A note on a general definition of the coefficient of determination." *Biometrika* 78: 691-692. DOI: 10.1093/biomet/78.3.691.

Olson, K., and R.M. Groves. 2012. "An examination of within-person variation in response propensity over the data collection field period." *Journal of Official Statistics* 28: 29-51.

Olson, K., J.D. Smyth, and H.M. Wood. 2012. "Does giving people their preferred survey mode actually increase survey participation rates? An experimental examination." *Public Opinion Quarterly* 76: 611 – 635. DOI: 10.1093/poq/nfs024.

Pepe, M.S. 2003. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.

Pickery, J., G. Loosveldt, and A. Carton. 2001. "The effects of interviewer and respondent characteristics on response behavior in panel surveys - a multilevel approach." *Sociological Methods and Research* 29: 509–523. DOI: 10.1177/0049124101029004004.

Plewis, I., S. Ketende, and L. Calderwood. 2012. "Assessing the accuracy of response propensities in longitudinal studies." *Survey Methodology* 38 (2): 167-171.

Potthoff, R.F., K.G. Manton, and M.A. Woodbury. 1993. "Correcting for nonavailability bias in surveys by weighting based on number of callbacks." *Journal of the American Statistical Association, Applications and Case Studies* 88 (424): 1197-1207. DOI: 10.1080/01621459.1993.10476399.

Schwarz, G. 1978. "Estimating the dimension of a model." *The annals of statistics* 6(2): 461-464. DOI: 10.2307/2958889.

Sinibaldi, J. 2014. “Using call-level interviewer observations to improve response propensity models.” In *Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications* (*PhD Thesis)*. München: Ludwig-Maximilian-Universität.

Sinibaldi, J., G.B. Durrant, and F. Kreuter. 2013. “Evaluating the measurement error of interviewer observed paradata.” *Public Opinion Quarterly, Special issue: Topics in Survey Measurement and Public Opinion* 77 (1): 173-193. DOI: 10.1093/poq/nfs062.

Sinibaldi, J., M. Trappmann, and F. Kreuter. 2014. “Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations?” *Public Opinion Quarterly* 78(2): 440-473. DOI: 10.1093/poq/nfu003.

Storr, C. L., H. Zhou, K.Y. Liang, and J.C. Anthony. 2004. “Empirically derived latent classes of tobacco dependence syndromes observed in recent-onset tobacco smokers: epidemiological evidence from a national probability sample survey.” *Nicotine and Tobacco Research* 6(3): 533-545. DOI: 10.1080/14622200410001696493.

Wagner, J. 2013a. “Adaptive contact strategies in telephone and face-to-face surveys.” *Survey Research Methods* 7: 45-55. DOI: http://dx.doi.org/10.18148/srm/2013.v7i1.5037.

Wagner, J. 2013b. “Using paradata-driven models to improve contact rates in telephone and face-to-face surveys.” In *Improving surveys with paradata: Analytic use of process information,* edited by F. Kreuter, 145-170. New Jersey: Wiley and Sons.

Watson, N., and M. Wooden. 2009. "Identifying factors affecting longitudinal survey response." In *Methodology of longitudinal surveys*, edited by P.Lynn, 157-182. New York: John Wiley and Sons.

West, B. T. and R.M. Groves. 2013. “A propensity-adjusted interviewer performance indicator.” *Public Opinion Quarterly 77*: 352-374. DOI: 10.1093/poq/nft002.

White, H. 1980. “A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.” *Econometrica* 48: 817–830. DOI: 10.2307/1912934.

White, H. 1984. *Asymptotic theory for econometricians*. Orlando, FL: Academic Press.

White, H. 1994. *Estimation, inference and specification analysis*. New York: Cambridge University Press.

Tables

Table 1: Distributions of the three response variables in the final analysis sample (total 10,630 households).

Variables with categories	Frequencies	Percentages
Length		
Short sequence (up to 6 calls)	6704	63.1
Long sequence (7-30 calls)	3926	36.9
Final outcome		
No single interview in a sequence after call 3	3110	29.3
At least one interview in a sequence after call 3	7520	70.7
Combined response		
Short successful	5304	49.9
Short unsuccessful	1400	13.2
Long successful	2216	20.8
Long unsuccessful	1710	16.1

Table 2: Different evaluation criteria to allow comparisons of the three types of models for length, final response outcome and the combined dependent variable of length and final response outcome (Nagelkerke's pseudo R^2, the overall percentage of correctly classified households provided by the classification tables and the Area under the Curve (AUC) from the Receiver Operating Curves (ROC)).

	Model	Length			Outcome			Combined	
		pseudo R^2	Classification Table	AUC	pseudo R^2	Classification Table	AUC	Pseudo R^2	Classification Table
1	Just geographic and design variables from W1	0.019	63.1	0.570	0.025	70.7	0.582	0.033	49.9
2	+ survey W1	0.055	63.6	0.618	0.055	71.0	0.622	0.081	50.1
3	+ interviewer observation W1	0.060	63.8	0.624	0.062	71.0	0.629	0.092	50.3
4	+ call record W1	0.080	64.5	0.643	0.072	71.1	0.640	0.113	50.4
4b	Model 3 +latent classes +length of sequence	0.078	64.5	0.642	0.067	71.1	0.635	0.108	50.2
5	Model 4 +interviewer observations W2 + HH split indicator	0.090	65.0	0.653	0.128	72.3	0.688	0.159	51.3
6	+change between interviewer observation W1 and W2 indicators	0.095	65.3	0.657	0.141	72.6	0.698	0.171	51.5
7	Model 6+ all call record (all 3 calls) W2 (without call outcomes)	0.110	66.0	0.668	0.181	73.7	0.724	0.219	52.4
8	+ call outcomes for 3 calls W2	0.242	69.3	0.751	0.270	75.6	0.777	0.362	56.0

Note: HH = household; W1 = Wave 1; W2 = Wave 2.

Table 3: Results of the classification table showing the percentage of correctly classified (discriminated) households by categories of the two binary and the multinomial dependent variable for each of the 8 modelling stages considered. (Column percentages shown, i.e. percentage of those households which were estimated correctly out of the total observed in the group.) (For the two binary outcomes these are sensitivity $P(\hat{y}_i = 1|y_i = 1)$ and specificity $P(\hat{y}_i = 0|y_i = 0)$, and for the multinomial model it is $P(\hat{y}_i = s|y_i = s)$, for $s = 1,2,3,4$).

	Length		**Final Outcome**		**Combined Length and Outcome**			
Model	**Short**	**Long**	**Successful**	**Un successful**	**Short Successful (n=5304)**	**Short Unsuccessful (n=1400)**	**Long Successful (n=2216)**	**Long Unsuccessful (n=1710)**
1	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
2	93.2%	13.1%	98.6%	4.3%	98.6%	0.3%	0.1%	5.5%
3	92.2%	15.2%	97.9%	5.8%	97.6%	0.5%	1.4%	7.8%
4	89.4%	22.1%	97.3%	7.7%	95.0%	0.8%	4.9%	11.5%
4b	89.7%	21.6%	97.6%	6.8%	95.2%	0.6%	4.8%	10.1%
5	88.7%	24.6%	94.9%	17.8%	93.2%	3.1%	5.6%	20.4%
6	88.5%	25.6%	94.2%	20.4%	92.7%	3.8%	5.5%	22.5%
7	87.5%	29.3%	93.3%	26.4%	90.4%	14.1%	8.9%	22.0%
8	80.8%	49.8%	91.6%	36.7%	84.4%	23.1%	28.1%	31.1%

Table 4: Results of the classification table showing the percentage of correctly predicted households by categories of the two binary and the multinomial dependent variable for each of the 8 modelling stages considered. (Row percentages shown, i.e. percentage of those households which were observed correctly out of the total estimated in the group). (For the two binary outcomes these are the positive ($P(y_i = 1|\hat{y}_i = 1)$) and negative predicted values ($P(y_i = 0|\hat{y}_i = 0)$), and for the multinomial model it is $P(y_i = s|\hat{y}_i = s)$, for $s = 1,2,3,4$).

	Length		**Final Outcome**		**Combined Length and Outcome**			
Model	**Short**	**Long**	**Successful**	**Un successful**	**Short Successful (n=5304)**	**Short Unsuccessful (n=1400)**	**Long Successful (n=2216)**	**Long Unsuccessful (n=1710)**
1	63.1%	0.0%	70.7%	0.0%	49.9%	0.0%	0.0%	0.0%
2	64.7%	53.1%	71.4%	55.8%	50.5%	66.7%	37.5%	34.6%
3	65.0%	53.4%	71.5%	53.7%	50.9%	46.7%	37.2%	37.2%
4	66.2%	54.9%	71.8%	54.3%	51.8%	39.3%	32.6%	36.6%
4b	66.1%	55.1%	71.7%	54.4%	51.6%	42.1%	31.5%	34.6%
5	66.8%	56.1%	73.6%	54.3%	53.3%	43.0%	38.3%	37.2%
6	67.0%	56.5%	74.1%	59.0%	53.7%	37.9%	35.0%	39.1%
7	67.9%	57.8%	75.4%	61.9%	54.8%	48.3%	36.8%	40.4%
8	73.3%	60.3%	77.8%	64.5%	62.1%	48.6%	42.2%	41.1%

Table 5: Complete classification table for the multinomial model (dependent variable is combined length and outcome) for Model 8: (A) column percentages (percentages predicted out of the total observed in the category) reflecting sensitivity of modelling long unsuccessful calls and (B) row percentages (percentages of households observed in the group out of the total predicted in the category) reflecting positive predictive values.

			Observed			
			Short Successful (n=5304)	**Short Unsuccessful (n=1400)**	**Long Successful (n=2216)**	**Long Unsuccessful (n=1710)**
A (Discrimination)		**Short Successful**	84.4%	50.1%	58.0%	43.5%
	Predicted	**Short Unsuccessful**	2.8%	23.1%	2.7%	7.9%
		Long Successful	8.0%	9.2%	28.1%	17.6%
		Long Unsuccessful	4.9%	17.5%	11.2%	31.1%
B (Prediction)		**Short Successful**	62.1%	9.7%	17.8%	10.3%
	Predicted	**Short Unsuccessful**	22.3%	48.6%	8.8%	20.2%
		Long Successful	28.7%	8.7%	42.2%	20.4%
		Long Unsuccessful	20.1%	19.1%	19.3%	41.4%