This paper has been accepted for publication by the Journal of Survey Statistics and Methodology, 17 March 2015.

Modelling Final Outcome and Length of Call Sequence to Improve Efficiency in Interviewer Call Scheduling

Gabriele B. Durrant, Olga Maslovskaya and Peter W.F. Smith

Southampton Statistical Sciences Research Institute
University of Southampton
United Kingdom
(g.durrant@southampton.ac.uk)

Address for correspondence:
Gabriele Durrant
Southampton Statistical Sciences Research Institute
University of Southampton
SO17 1BJ, Southampton
United Kingdom
g.durrant@southampton.ac.uk

Abstract:

Survey practitioners are increasingly interested in how best to use paradata to improve data collection processes. One particular question is if it is possible to identify early on during fieldwork sample cases that may require a long time, and therefore a lot of financial and staff resources, until interviewing is completed. More specifically, we aim to identify cases with long unsuccessful call sequences. This paper models call record data predicting final call outcome and length of a call sequence. Separate binary and joint multinomial logistic models for the two outcomes are presented, accounting for the clustering of households within interviewers. Of particular interest is to identify explanatory variables that predict final outcome and length of a call sequence. The study uses data from Understanding Society, a large-scale UK longitudinal survey. The work has implications for responsive and adaptive survey designs. The results indicate that modelling outcome and length of a call sequence jointly improves the fit of the model. Outcomes of previous calls, in particular from the most recent call, are highly predictive. The timing of calls and interviewer observation variables, although significant in the models, only slightly improve the predictive power.

Key Words: survey non-response, interviewer call record data, paradata, 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.

Data Statement

This study uses Wave 1 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

For interviewer administered surveys many statistical agencies nowadays routinely collect call record data. Examples of such data are recordings of the day, time and outcome of each call or visit and any observations made about the person talked to at the call. For face-to-face surveys a range of interviewer observation variables, such as physical and social characteristics of the selected household and neighbourhood, may also be recorded. Researchers have become increasingly interested in how best to use and analyse such information. It is hoped that a better understanding of the calling patterns and the mechanisms leading to particular call sequences help improve data collection through reducing both costs and non-sampling errors. For statistical agencies, investing time and effort into repeated calls and follow-ups to a sample unit is very resource- and cost-intensive. It is therefore desirable to avoid long unsuccessful call sequences to improve efficiency of call scheduling. The aim is to identify cases prone to long and unsuccessful call sequences.

This paper presents models for call record data predicting final call outcome and final length of a call sequence early on in the data collection process, for example after the first, second or third call. Separate binary logistic models and a multinomial logistic model for the two outcomes combined are considered. The clustering of sample cases within interviewers is taken into account by using robust standard error estimation. To assess the models, a range of methods are employed, including the widely used R^2 statistic, classification tables and ROC (receiver operating characteristics) curves (Agresti 2013). Concepts from epidemiology are introduced here into the context of survey methodology including discrimination and prediction (Plewis, Ketende, and Calderwood 2012; Pepe 2003). A particular focus is the identification of explanatory variables that predict final outcome and/or length, especially those characterising

long unsuccessful call sequences. Further research questions that we aim to address in this study are: how can call record predictors best be incorporated into the model(s): as summary statistics or as individual outcomes?; how predictive are the models?; does their ability to predict improve once interviewer observation variables and more and more call records are available?

Past research mostly aimed to predict final non-response, often did not include paradata and relied on fully observed frame data or socio-demographic variables. However, such models have not been found to predict well the outcomes of calls (Groves and Couper 1996; Groves and Couper 1998; Bates, Dahlhamer, and Singer 2008). In recent years, researchers have explored the potentials of including newly available paradata, such as call record data and interviewer observation variables, in models for non-response outcomes with some success (Potthoff, Manton, and Woodbury 1993; Groves and Couper 1996; Sinibaldi, Durrant, and Kreuter 2013; Sinibaldi, Trappmann, and Kreuter 2014; Bates, Dahlhamer, and Singer 2008; Wagner 2013a and 2013b; Kreuter et al. 2010), but typically still with low predictive power (R^2 values between 3%-8%) (Olson, Smyth and Wood 2012; Olson and Groves 2012; West and Groves 2013). Also, the length of a call sequence to obtain a response has not been the focus and has been, if at all, used only as an explanatory variable in models. Furthermore, the use of discrete time event history analysis to predict the outcome at the next call rather than the final outcome has been advocated, exploiting call record data as explanatory variables (Groves and Heeringa 2006; Durrant, D'Arrigo, and Steele 2011 and 2013; Durrant, D'Arrigo, and Müller 2013; Sinibaldi 2014; Hanly 2014). However, the outcome over the next few calls may be of greater interest than the outcome of the next call only.

One novel aspect of this paper is to consider modelling *both* length and outcome of call sequences simultaneously. For comparison, separate and joint models for the two outcomes are

developed. The study uses data from a large-scale face-to-face longitudinal survey in the UK, Understanding Society. The research is motivated by findings from a recent sequence analysis of the same data (Durrant, Maslovskaya and Smith 2014), which highlights the importance of sequence length and outcome as two key characteristics of the large number of different sequences, supporting findings from earlier work (Kreuter and Kohler 2009). This analysis identified a significant number of long call sequences (up to 30 calls) for some households and a distinct grouping of sequences into short and long, successful and unsuccessful call sequences.

The research has implications for survey practice, in particular for adaptive and responsive survey designs. The method developed may be of particular benefit for longitudinal surveys, where the same or similar auxiliary variables are available for every wave. The models could then be used to predict final outcome and sequence length for future waves.

The remainder of the paper is structured as follows. First, the Understanding Society survey and the analysis sample are discussed. The methods section presents the regression models, the modelling strategy and the methods to compare and assess different models. Then, the results are presented. The paper concludes with a summary of the main findings, implications for survey practice, limitations and further work.

2. Data

2.1 The Understanding Society Survey

This paper uses call record data and interviewer observations from the first wave of the UK Understanding Society survey, which is the Household Longitudinal Study in the United Kingdom. The survey covers topics of health, work, education, income, family and social life to help understand the long term effects of social and economic change, as well as policy

interventions. The survey has a multi-stage sampling design with clustering and stratification. For further information on the design of the survey see Buck and McFall (2012). The study has many advantages over previously existing datasets in the UK by being exceptionally large and comprehensive. In particular, as part of the study a range of paradata were collected, including call record data and a wide range of interviewer observation variables. Also, only interviewers with above average experience and ability were selected for the study.

Data collection for each wave is scheduled across a 24-months period, with interviews taking place annually. Wave 1 data collection took place between January 2009 and March 2011. All Wave 1 interviews were carried out face-to-face in respondents' homes by trained interviewers using computer-assisted personal interviewing (CAPI). Households are therefore clustered within interviewers. All adult household members (age 16 and older) are asked to respond. In addition to individual interviews, a member of the household needs to respond to a household questionnaire. Interviewers have one month to contact households. A minimum of six calls are made at each sampled address before it is considered unproductive, but interviewers are encouraged to make further calls if possible (McFall 2012).

During the first visit to a household, interviewers collect a wide range of *interviewer observations* capturing characteristics about each household and surrounding neighbourhood. (see online Appendix Table A1 for wording of all variables considered). In addition, *call record data* are available, which capture information about each call, including outcome, date and time of each call. The outcome of a call in the Understanding Society survey is defined as non-contact, contact, appointment, interview, and 'any other status' (which includes ineligibles and refusals as coded by the survey agency). For this study, call record data, including final response outcome, are combined with interviewer observation data, variables about the study design and geographic

information, using a unique household identifier. All variables are available for both respondents and non-respondents.

2.2 Analysis sample and definition of explanatory variables

The analysis was initially carried out on all households from Wave 1, excluding cases from an Ethnic Minority Boost sample as rules for the selection of this sample differ from the main sample. A separate analysis could be undertaken for this group. This distinction is, however, not of interest here. Ineligibles are included in our analysis since it is a true outcome and survey agencies may be interested in mechanisms for identifying ineligibles as early as possible to reduce survey costs. This *initial analysis sample* contained 47,913 households (including both responding and non-responding households). First, we carried out preliminary work using this initial analysis sample, conducting separate analyses for all cases with at least 1, 2, 3 etc. calls. The aim was to evaluate after every call, how well we can predict final outcome and length as more information about each case becomes available. This preliminary analysis suggested that 3 calls may be sufficient to reach an acceptable level of predictability. To more formally assess this we explore a range of models to predict length and outcome. However, to be able to compare model fit and prediction across the different models -since measures of predictability are dependent on sample size- we have to restrict our final analysis sample for evaluation purposes only to all households from Wave 1 that received four or more calls (27,995 households), although in survey practice the models can be fitted to all the available cases. There are no missing cases in any 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, contained a very small amount of missing data, and such calls were excluded (311 cases). Since

we are interested in the usefulness of interviewer observation variables we condition on the availability of such variables (removing 2,015 cases, equivalent to 7.2%). The *final analysis sample*, including only cases with four or more calls, contains 25,358 households within 734 interviewers.

The explanatory variables used in the analysis to model call sequence length and outcome can be split into three main groups (for the full list of variables see the online Appendix, Table A1). The distributions of the explanatory variables are presented in the online Appendix (Table A2).

1.) *geographic information and design variables* (4 variables: urban/rural indicator, government office region, low density area for ethnic minorities and month of household issue);
2.) *interviewer observation variables* (14 variables, e.g. indicators of entry barriers, 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);
3.) *call record variables* (20 variables, e.g. time of day, day of week, call outcome; also derived variables including time between calls, number of previous non-contacts, contacts, appointments and broken appointments).

3. Methods

To analyse length and call outcome a range of different specifications of the dependent variables are considered. To correctly control for the clustering of households within interviewers robust standard error estimation is employed (Huber 1967; White 1980, 1984 and 1994). As an alternative modelling approach, multilevel models could have been employed also

allowing for the nesting of households within interviewers. However, since interviewer effects are not of substantive interest here (e.g. we are not interested in explaining effects of contextual or interviewer-level variables), a multilevel modelling approach is not necessary. To assess predictability of different models we computed R^2 statistics, classification tables and ROC curves.

3.1 Modelling call sequence length and outcome

We first define the dependent variables in our models. Given our research questions and the design of the survey, we are interested in modelling short and long sequences, successful and unsuccessful call outcomes and the combination of both. Due to the survey protocol requirement of conducting a minimum of 6 call attempts, if contact was not established, it follows naturally to define short and long sequences at the cut-off of 6 calls for this study. A successful call outcome is defined as achieving at least one interview per household.

Since other cut-offs and other non-binary specifications of the dependent variable 'length of call sequence' are also possible, we carried out a detailed sensitivity analysis, a.) to see how sensitive the results are to the choice of cut-off and b.) to allow for other specifications. We considered three binary specifications (cut-offs at 5, 6 and 7 calls), applying binary logistic models, and two categorical specifications (3 categories with cut-offs at 6 and 9 calls; 4 categories with cut-offs at 6, 8 and 11 calls), applying both multinomial and ordinal regression to each categorisation. To allow for the integer nature of the variable we also modelled it as a count variable (although it may be argued that predicting the exact call number required may be of less importance to survey practice than knowing the rough order of further calls). Poisson regression was applied with robust standard errors to control for overdispersion (Cameron and Trivedi, 1998; Long and Freese, 2006). Comparing the different models using the R^2 statistic, the binary

and Poisson models performed best. Across all models only very small differences in the significance and direction of effects of the explanatory variables were found. The key findings were not sensitive to the particular cut-offs chosen. The binary model with the cut-off at 6 calls only differed from the Poisson model in the significance of two coefficients. In particular, across all models all key explanatory variables (such as outcome of previous calls) showed consistent results, ensuring the overall conclusions in this paper to be robust to the particular model specification.

To summarise, in this paper we focus on the following three response variables and models:

1.) The dependent variable '*length of call sequence'* is coded

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

where y_i denotes a binary response variable of household i.

2.) The dependent variable *'final outcome of call sequence'* is coded

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

3.) For the dependent *variable combining both length and outcome* y_i is coded

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

Given our analysis sample with the aim of allowing for comparison of models, all call outcomes are with reference to after 3 calls.

For the two binary dependent variables logistic models are employed. 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):

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

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

For the combined variable multinomial modelling is employed. 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

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

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

3.2 Modelling strategy and comparison of 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 the model. For data preparation and analysis we used SPSS version 20 and STATA 12. STATA can estimate robust standard errors to control for the non-independence of observations (***ro*** function; Long and Freese 2006). This approach is used here to account for all levels of clustering. Also, all models include as a minimum geographical stratification variables to account for the primary stratification. To simplify the interpretation of

the modelling results predicted probabilities are obtained for specific variables, holding all other variables constant at their means.

3.2.1 Modelling strategy and evaluation

Explanatory variables, together with selected interaction effects, are considered step by step, depending on the type of data available after each call. (a. only geographic and design variables; b. interviewer observations, information about the timing of the first call attempt, then timing of second and third call (including time between calls); c. information about the outcome for the first three calls one by one and as combinations, either as raw outcomes (i.e. outcome of first, second and third call, interactions between outcomes) or as summary information (i.e. number of non-contacts, contacts, appointment and interviews across the first three calls)). Allowing for interaction effects between outcomes accounts for the sequence of outcomes as a whole rather than simply as individual outcomes. For example, a non-contact call after an appointment should be interpreted differently from a non-contact after a previous non-contact. The former may reflect an indirect refusal, whereas the latter indicates a longer period of absence.

To decide on the final model a forward stepwise procedure is used. To test the sensitivity of the final variable selection two other variable selection techniques are also employed (backward stepwise and enter method). For the two logistic models all methods led to the same results. For the multinomial model they only differed in two variables and then on effects which were only marginally significant. As discussed in section 3.1 we also fitted different types of models (logistic, multinomial, ordinal and Poisson) and they also led to the inclusion of the same or very similar variables, ensuring the stability of the variable selection. For model checking purposes the

distribution of the standardised residuals are investigated, as well as boxplots of the standardised residuals against the explanatory variables in the model. Checking for outliers (standardised residuals greater than 1.96 in absolute value) we found 2.7% of outliers in the model for length and 3.1% in the model for outcome, which is acceptable. Only selected models will be discussed further in this paper. It should be noted that ideally survey researchers would want to control for household size in the models. Here, this was not directly possible since this variable is only observed for responding households. Although there is no direct measure of household size from the interviewer observation variables in Understanding Society, a number of proxy measures are used, such as type of accommodation, floor level and presence of children.

3.2.2 Comparison of model performance

To compare the different models and to assess the quality of model predictions and model fits, we employ the R^2 statistic (Nagelkerke R^2) (Field 2009; Long and Freese 2006), classification tables and ROC curves (Agresti, 2013). Although the R^2 statistic is widely used to assess the prediction performance of non-response models (Groves and Couper 1998; Bates, Dahlhamer, and Singer 2008), it is designed to assess the overall fit of the model and therefore does not distinguish between the accuracy of the model for respondents and non-respondents separately (Plewis, Ketende, and Calderwood 2012). To achieve this we borrow ideas from concepts widely used in epidemiology. Accuracy may be determined by two related concepts: *discrimination* and *prediction* (Pepe 2003). Discrimination refers to the conditional probability that a household is predicted to be a respondent (non-respondent) given that a household is indeed a respondent (non-respondent). Prediction refers to the conditional probability of being a respondent (non-respondent) given a household is predicted to be a respondent (non-respondent)

(Plewis, Ketende, and Calderwood 2012). To evaluate both concepts classification tables are derived cross-classifying the observed binary response with a prediction of whether $y_i=0$ or $y_i=1$ (Agresti 2013). The prediction for an observation will be obtained depending on a cut-off π_0. The prediction for observation i is $\hat{y}_i=1$ if $\hat{\pi}_i>\pi_0$, and $\hat{y}_i=0$ if $\hat{\pi}_i\leq\pi_0$, where $\hat{\pi}_i$, denotes the predicted probability from the model. The discrimination power may be summarised by *sensitivity*=$P(\hat{y}_i=1|y_i=1)$ (or true positive fraction) and *specificity*= $P(\hat{y}_i=0|y_i=0)$ (1 minus the false positive fraction) (Agresti 2013; Plewis, Ketende, and Calderwood 2012; Altman 1991). As an overall summary measure of model performance, the percentage of observations correctly classified may be used (i.e. a summary of the diagonal of the classification table), which is a weighted average of sensitivity and specificity:

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

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

The concept of prediction is useful, since it tackles the problem from the other direction. In practice, survey researchers would not know who will be a respondent or a non-respondent before the end of data collection. This means we are interested in the probability of the prediction being correct whether the household is in reality a respondent or a non-respondent. These are measured by the *positive predictive value*, $P(y_i=1|\hat{y}_i=1)$, and the *negative predictive value*, $P(y_i=0|\hat{y}_i=0)$. The concepts of classification table, discrimination and prediction can be extended to the multinomial case, allowing for several groups of misclassifications. For a 4 category variable this results in a 4×4 classification table, allowing for 4 correctly and 12 incorrectly classified groups. In this analysis we are particularly interested in predicting non-response and in discriminating between respondents and non-respondents. In the results section

we therefore refer to sensitivity and positive predicted values with respect to modelling long and unsuccessful call sequences. In our analysis we use the default option of π_0=0.50 for the binary and π_0=0.25 for the multinomial case, although in practice different values can be explored to optimise the balance between discrimination and prediction (Plewis, Ketende, and Calderwood 2012).

A potential restriction of the classification table and the evaluation of prediction and discrimination is their dependency on a cut-off value π_0. A ROC curve (Agresti, 2013; Plewis et al. 2012; Krzanowski and Hand, 2009) summarizes predictive power for all possible values of π_0, by plotting sensitivity as a function of (1-specificity) across all values for π_0. The ROC curve usually has a concave shape connecting the points (0.0) and (1,1), which results from π_0 ranging from 0 to 1. The higher the sensitivity the higher is the predictive power, given a particular specificity. This implies that the greater the area under the curve (AUC) (i.e. between the curve and the horizontal-axis) the greater the predictive power. The AUC values can range from 1 (perfect discrimination) to 0.5 (no discrimination), i.e. the area below the diagonal, implying predictions to be no better than random guesses. For comparing the predictive power across the different models and across cut-off values π_0, we also provide the AUC values.

4. Results

The distributions of the three response variables are presented in Table 1. A range of models with increasing amount of explanatory variables are fitted to each of the three dependent variables. Table 2 presents pseudo-R^2 coefficients and the results from the classification tables of the percentage of correctly classified households for eight selected models.

Table 1: Distributions of the three response variables in the final analysis sample (25,358 households).

Variables with categories	Frequencies	Percentages
Length		
Short sequence (up to 6 calls)	12353	48.7
Long sequence (7-30 calls)	13005	51.3
Final outcome		
No single interview in a sequence	13565	53.5
At least one interview in a sequence	11793	46.5
Combined response		
Short unsuccessful	4962	19.6
Short successful	7391	29.1
Long unsuccessful	8603	33.9
Long successful	4402	17.4

To be able to interpret the results from the classification tables correctly let us consider predicting households to be in the most frequently observed category of the response variables. For the variable length we would expect about 51% (see Table 1) of cases to be correctly predicted, for the final outcome variable about 54%, and for the multinomial case about 34%. So to do better than predicting according to the marginal distribution of the dependent variables, we aim to find classification values above 51%, 54% and 34% respectively. Table 2 suggests that the predictive power of models only including geographic location and design variables (Model 1) is very low with pseudo-R^2 coefficients all below 3% and only about 55% of cases being correctly classified for both binary models and 36% for the multinomial model. Similarly, summarising predictive power, the area under the ROC curve (AUC) is only about 0.55 for the two binary models, which is not much better than chance. The introduction of interviewer observations to the models improves the predictability of the models in relative terms. The R^2 value doubles for length, triples for the multinomial model and quadruples for the final outcome variable. However, in absolute terms it is still below 9% across all three models. The classification of cases improves by 2.8-3.6 percentage points. The AUC values for the binary

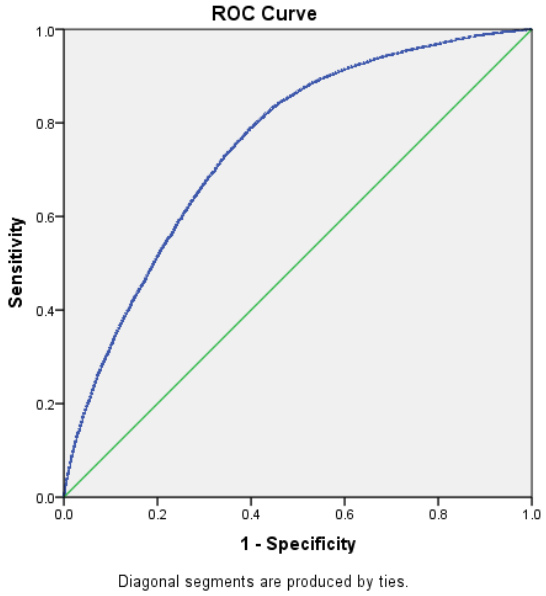
models increase to just above 0.60. Although there is a relatively low improvement in prediction performance, it should be noted that the majority of interviewer observation variables are found to be highly significant across all models, stressing the importance of such variables. The introduction of the call record information in the form of timings of calls slightly improves the models' predictions (Models 3 and 4). The R^2 is now around 10% for the binary and 18% for the multinomial model with the classification tables around 61% and 43% respectively. Including information about call outcomes (Models 6-8) substantially improves the models' prediction. The R^2 values increase to around 25% for the binary case and to even 37% for the multinomial case. The classification table has increased to about 70% for the two binary and 52% for the multinomial cases, reflecting an improvement in the percentage of correctly allocated cases by 37%, 25% and 53% in comparison to predictions based on the distribution of the dependent variables. Similarly, the AUC values have now increased to around 75%, which is significantly different from 0.5 ($p=0.000$) meaning that the models classify groups significantly better than chance. The values are higher than those reported for other nonresponse models (Plewis et al. 2012), implying that discrimination between nonrespondents and respondents is good. The ROC curves for Model 8 (length and outcome) are given in Figure 1. Comparing the three models including outcome(s) of previous calls, Model 8, which includes raw outcomes for all three calls, performs slightly better with regards to prediction than one outcome on its own (Model 6) or summaries of separate call outcomes (Model 7). As one may expect, the outcome of the last call (here call 3) is the key variable when predicting final outcome and length.

Table 2: Comparison of different models for length, (final) outcome and the combined dependent variable of length and outcome (Nagelkerke R^2, classification table, i.e. the percentage correctly classified households, and AUC (Area under the ROC curve) values).

	Model	Length			(Final) Outcome			Combined Outcome	
		Nagelkerke R^2	Classification table	AUC (ROC curve)	Nagelkerke R^2	Classification table	AUC (ROC curve)	Nagelkerke R^2	Classification table
1	**Just geographic**	0.027	55.5%	0.579	0.013	54.6%	0.557	0.029	36.2%
2	**1+interviewer observations**	0.062	58.7%	0.621	0.053	58.2%	0.609	0.085	39.0%
3	**2+call record for call 1 including call outcome**	0.078	59.9%	0.636	0.065	59.4%	0.624	0.115	40.3%
4	**2+call record for calls 1 and 2 including call outcomes**	0.121	61.9%	0.670	0.098	61.1%	0.655	0.184	43.2%
5	**2+call record for calls 1-3 without call outcomes**	0.110	61.1%	0.660	0.105	61.2%	0.656	0.185	42.9%
6	**5+call outcome for call 3**	0.248	69.1%	0.748	0.236	67.4%	0.739	0.360	51.2%
7	**5+4 sums of call outcomes across the calls 1-3**	0.224	68.4%	0.738	0.209	66.9%	0.730	0.332	50.4%
8	**5+call outcomes for calls 1-3**	0.258	69.7%	0.753	0.242	67.8%	0.744	0.371	51.7%

Figure 1: ROC (Receiver operating characteristics) curves for Model 8 modelling a.) call sequence length and b.) final outcome

a.) ROC curve for Model 8 modelling call sequence length



b.) ROC curve for Model 8 modelling final outcome

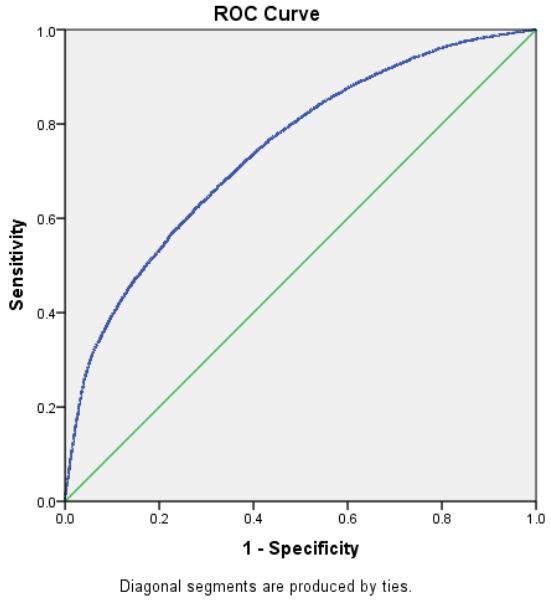


Table 2 suggests that the best models for prediction is obtained in Model 8 which contains variables from the geographic and design group, interviewer observations and call record data including raw call outcomes for the first three calls and timing of calls. Also, modelling outcome and length of call sequence jointly improves the fit of the model in comparison to the two separate models for either length or final outcome based on the R^2 value (highest R^2 value for the combined model) and the classification table (53% improvement in the percentage of correctly allocated cases for the combined model when compared to the predictions based on the distributions of the dependent variables; for the two separate models these are much lower with 37% and 25% respectively).

Table 3: Results of the classification table showing the percentage of correctly classified households by categories of the multinomial dependent variable (combined length and outcome) 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).

Model	Short Unsuccessful (n=4962)	Short Successful (n=7391)	Long Unsuccessful (n=8603)	Long Successful (n=4402)
1	0.0%	43.2%	69.6%	0.0%
2	6.5%	52.8%	65.9%	0.1%
3	20.4%	49.8%	64.2%	0.1%
4	28.9%	49.7%	67.7%	0.5%
5	31.1%	51.6%	63.9%	0.4%
6	44.3%	50.2%	79.5%	5.2%
7	42.2%	54.4%	75.3%	3.9%
8	45.1%	51.0%	79.5%	5.6%

In the non-response literature researchers are interested in predicting the non-respondents correctly. More specifically for this application, we are interested in predicting households with no interview and long call sequences early on in the data collection process. Table 3 shows the percentage of correctly classified households by the categories of the multinomial dependent variable for each of the 8 modelling stages considered. Across all stages the percentage of

correctly classified households with long unsuccessful call sequences (around 64-80%) is quite high. For Model 8 this value increases to 80%, meaning that of the cases that have long unsuccessful call sequences 80% are correctly classified by the model as being indeed in this category.

Table 4: 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 Unsuccessful (n=4962)	**Short Successful (n=7391)**	**Long Unsuccessful (n=8603)**	**Long Successful (n=4402)**
A	**Predicted**	**Short Unsuccessful**	45.1%	6.8%	9.8%	5.2%
		Short Successful	13.4%	51.0%	8.8%	16.8%
		Long Unsuccessful	40.6%	39.8%	79.5%	72.4%
		Long Successful	0.9%	2.4%	1.9%	5.6%
B	**Predicted**	**Short Unsuccessful**	58.7%	13.1%	22.1%	6.1%
		Short Successful	11.2%	63.6%	12.8%	12.4%
		Long Unsuccessful	13.4%	19.6%	45.7%	21.3%
		Long Successful	7.4%	28.3%	25.2%	39.1%

Table 4 breaks this down further for the multinomial model for Model 8, showing marginal summaries of the classification table. The upper panel (A) indicates sensitivity (for a model for long unsuccessful calls) and the lower panel (B) the positive predictive values. For example, we can see that of the cases that are predicted to have long unsuccessful call sequences (panel B) indeed 46% belong correctly to this category, 13.4% and 19.6% would be wrongly classified and belong in reality to the short successful and short unsuccessful groups respectively.

Misclassification to short sequences would not have in practice negative implications since six calls might be made anyway. Of the cases predicted to have long unsuccessful call sequences 21% turn out to have long successful calls.

Let us now turn to the discussion of the effects of different explanatory variables on the three outcomes. A number of variables are found to be consistently significant across all models. For example, for length these are months of household issue, urban/rural, region, low density ethnic minority area, some interviewer observation variables (accommodation, floor, car/van, child, unkempt garden, relative condition of property) and some call record variables (time of day, time between calls and call outcome variables). Consistently significant predictors for final outcome are region, interviewer observation variables (floor, car/van, child, relative condition of property), time of day (call 1), time between calls and, as might be expected, all call outcome variables, but, unlike in the models for length, they exclude urban/rural, unkempt garden and low density area. Consistently significant predictors for the combined response models are similar to the model for length alone including geographic and design variables (months, urban/rural, region), interviewer observation variables (locked gates, accommodation, floor, car/van, child, unkempt garden, relative condition of property) and call record variables (all time of day variables, time between calls and call outcome variables). Interestingly, day of the week is not significant in most models and not in any of the final models.

The results of the final models for all three response variables from Model 8 are discussed below. Table 5 presents estimated regression coefficients together with odds ratios from the two binary logistic models (Model 8). The results suggest that the odds of having a short call sequence are higher for properties with 2 or less floors (1.216-2.247 times), for households which definitely do not have a car (1.947 times), are unlikely to have children or do

not have children (1.193 and 1.232 times) and if there is no unkempt garden (1.225 times). The odds are lower in town houses (terrace houses) (1.230 times) and flats (1.344 times) when compared to detached houses, also lower (1.107 times) when properties are worse than others in the neighbourhood, and lower when the calls are made in the evening compared to morning calls. (It should be noted that we cannot interpret this as a causal effect since calling times are not allocated randomly but are merely determined by the interviewer). The model suggests a positive association for time between calls: the longer the time between calls, the higher the probability of a short sequence. Shorter sequences are also more likely when previous call outcomes are a contact, an appointment, any other status or an interview when compared to non-contact. There is a marked monotone increase in the effect of the call outcome variables across the three calls, indicating that although the outcome of each call is significant, it may be the most recent call that has the highest influence, rather than, for example, the first call. As one might expect, an appointment or interview at the third call increases the odds of a short sequence by between 7 to 10 times compared to having a non-contact at this call attempt.

Table 5: Estimated coefficients (selected†) for the two logistic regression models for length and (final) outcome including geographic and design variables, interviewer observation variables and call record variables comprising timing and outcome of calls (Model 8) (The full model is provided in the Appendix, Table A3).

	Model for Length			Model for (final) outcome		
Variable	**B**	**Robust SE**	**OR**	**β**	**Robust SE**	**OR**
Interviewer observations variables						
Accommodation						
Detached house/bungalow (ref)	0.000		1.000			
Semi-detached house/bungalow	-0.062	0.042	0.940			
Town house (terrace house)	-0.207	0.046	0.813***			
Flat	-0.296	0.061	0.744***			
Rented room	0.257	0.184	1.293			
Floor						
0 floors	0.809	0.236	2.247**	0.806	0.213	2.238***
1 floor	0.318	0.078	1.374***	0.430	0.073	1.538***
2 floors	0.196	0.075	1.216**	0.402	0.065	1.495***
3 floors	0.051	0.079	1.052	0.413	0.077	1.511***
4 floors and above (ref)	0.000		1.000	0.000		1.000
Car/van						
Definitely has a car/van (ref)	0.000		1.000	0.000		1.000
Likely	-0.039	0.041	0.962	-0.189	0.039	0.827***
Unlikely	0.153	0.078	1.165	-0.452	0.075	0.636***
Definitely does not have a car/van	0.666	0.095	1.947***	0.654	0.095	1.924***
Cannot tell from observation	-0.137	0.041	0.872**	-0.400	0.036	0.670***
Child						
Definitely has a child/children aged under 10 (ref)	0.000		1.000	0.000		1.000
Likely	0.013	0.075	1.013	-0.099	0.075	0.906
Unlikely	0.177	0.068	1.193**	-0.272	0.065	0.762***
Definitely does not have a child/children aged under 10	0.209	0.068	1.232**	-0.163	0.066	0.849*
Cannot tell from observation	0.058	0.062	1.060	-0.312	0.061	0.732***
Unkempt garden						
Yes (ref)	0.000		1.000			
No	0.203	0.056	1.225***			
No obvious garden	0.113	0.062	1.120			
Relative conditions of the address to other residential properties						
Better (ref)	0.000		1.000	0.000		1.000

About the same	-0.047	0.054	0.954	-0.242	0.053	0.785***
Worse	-0.202	0.078	0.903*	-0.440	0.074	0.644***
Unable to obtain information	0.008	0.239	1.008	-1.248	0.260	0.287***
Call Record Variables						
Time of day call 1						
Morning (0.00-12.00) (ref)				0.000		1.000
Afternoon (12.00-17.00)				0.010	0.036	1.011
Evening (17.00-24.00)				-0.130	0.050	0.878**
Time of day call 2						
Morning (0.00-12.00) (ref)	0.000		1.000			
Afternoon (12.00-17.00)	-0.064	0.038	0.937			
Evening (17.00-24.00)	-0.102	0.043	0.903*			
Time of day call 3						
Morning (0.00-12.00) (ref)						
Afternoon (12.00-17.00)	-0.015	0.039	0.985			
Evening (17.00-24.00)	-0.094	0.041	0.912*			
Time between call 1 and call 2	0.026	0.002	1.026***	-0.026	0.002	0.974***
Time between call 2 and call 3	0.030	0.001	1.030***	-0.028	0.002	0.973***
Call 1 outcome						
No contact (ref)	0.000		1.000	0.000		1.000
Contact made	0.081	0.037	1.085*	-0.098	0.037	0.907**
Appointment made	0.025	0.069	1.026	0.242	0.065	1.274***
Any other status	0.124	0.095	1.133	-0.035	0.096	0.966
Interview done	1.022	0.240	2.780***	0.372	0.207	1.450
Call 2 outcome						
No contact (ref)	0.000		1.000	0.000		1.000
Contact made	0.095	0.038	1.099*	-0.062	0.037	0.940
Appointment made	0.205	0.063	1.228**	0.458	0.062	1.581***
Any other status	0.293	0.088	1.340**	-0.024	0.092	0.977
Interview done	1.539	0.143	4.662***	0.517	0.116	1.677***
Call 3 outcome						
No contact (ref)	0.000		1.000	0.000		1.000
Contact made	0.499	0.037	1.645***	-0.148	0.037	0.862***
Appointment made	2.000	0.048	7.389***	2.024	0.053	7.568***
Any other status	1.227	0.065	3.410***	-1.185	0.074	0.305***
Interview done	2.352	0.110	10.511***	0.860	0.087	2,364***

***p<0.001; **p<0.01; *p<0.05; ref – reference category

† The models also include a constant and control for geographic and design variables.

In the model for final outcome some of the same variables are significant as in the model for length, however, sometimes with the opposite effect. For example, households with no children are significantly less likely to respond and a longer time in between calls is associated with a reduced likelihood of response. A contact in any of the previous calls is associated with a higher likelihood of non-response as is any other status. Appointment made and any interviewing performed increases the likelihood of a successful call sequence as one would expect. The effect of number of floors and the presence of cars/vans are in the same direction as for length of call sequence.

Table 6 presents estimated coefficients for the multinomial model. To ease interpretation predicted probabilities are computed for selected explanatory variables holding constant all other variables at their mean value (Figure 2; also provided in color in the online appendix). Variables from the interviewer observations (accommodation type) and call record variables (time of day of first call, third call outcome) are chosen. Figure 2a suggests that the probability of having long unsuccessful sequences is the highest in flats compared to other types of accommodation with short successful calls highest in detached houses. Figure 2b suggests that conducting the first call in the evening is associated with a higher probability of a long unsuccessful call sequence. (Researchers need to be cautious in interpreting this as a causal effect since interviewers may have reasons for calling in the evening, even at the first call). Figure 2c indicates a marked difference between call outcomes, with the likelihood for short successful call sequences (getting an interview in calls 4-6) clearly highest if the third call is an appointment or an interview. Long unsuccessful call sequences are associated with non-contacts and to a slightly lesser extent with contact. The likelihood of having a short unsuccessful sequence is clearly highest when the third call attempt is recorded as being ‘any other status’.

Table 6: Estimated coefficients (selected†) for the multinomial regression model for the combined dependent variable (length and final outcome) including interviewer observation variables and call record variables comprising timing and outcome of the calls (Model 8). (Full model results are given in the Appendix, Table A4).

	Short Unsuccessful		Short Successful		Long Successful	
Variable	**β**	**Robust SE**	**β**	**Robust SE**	**β**	**Robust SE**
Interviewer observation variables (selected)						
Accommodation						
Detach. house (ref)	0.000		0.000		0.000	
Semi-detached	-0.073	0.059	-0.072	0.054	-0.011	0.060
Town House	-0.320***	0.066	-0.101	0.059	0.089	0.064
Flat	-0.604***	0.086	-0.147	0.077	-0.111	0.082
Rented room	0.363	0.242	0.268	0.237	0.163	0.266
Car/van						
Definitely has a car/van (ref)	0.000		0.000		0.000	
Likely	0.067	0.058	-0.206***	0.052	-0.147**	0.056
Unlikely	0.329**	0.105	-0.271**	0.101	-0.500***	0.110
Definitely does not have a car/van	0.838***	0.157	1.101***	0.133	-0.746***	0.138
Cannot tell from observation	0.043	0.058	-0.468***	0.053	0.328***	0.055
Child						
Definitely has a child/children aged under 10 (ref)	0.000		0.000		0.000	
Likely	-0.056	0.115	-0.055	0.096	-0.162	0.098
Unlikely	0.237*	0.099	-0.068	0.084	-0.330***	0.086
Definitely does not have a child/children aged under 10	0.314**	0.102	0.029	0.086	-0.164	0.087
Cannot tell from observation	0.095	0.079	-0.185*	0.079	-0.370***	0.080
Call Record Variables (selected)						
Time of day call 3						
Morning (ref)	0.000		0.000		0.000	
Afternoon	-0.025	0.054	0.028	0.050	0.058	0.053
Evening	-0.127*	0.057	-0.083	0.052	-0.034	0.055
Time between call 1 and call 2	0.034***	0.002	-0.002	0.002	-0.024***	0.003
Time between call 2 and call 3	0.038***	0.002	-0.001	0.002	-0.028***	0.003
Call 3 outcome						
No contact (ref)	0.000		0.000		0.000	
Contact made	0.618***	0.050	0.273***	0.049	-0.203***	0.052
Appointment made	0.187	0.102	2.683***	0.063	0.568***	0.080
Any other status	1.780***	0.077	-0.125	0.102	-0.661***	0.119
Interview done	2.344***	0.169	2.704***	0.157	0.556**	0.195

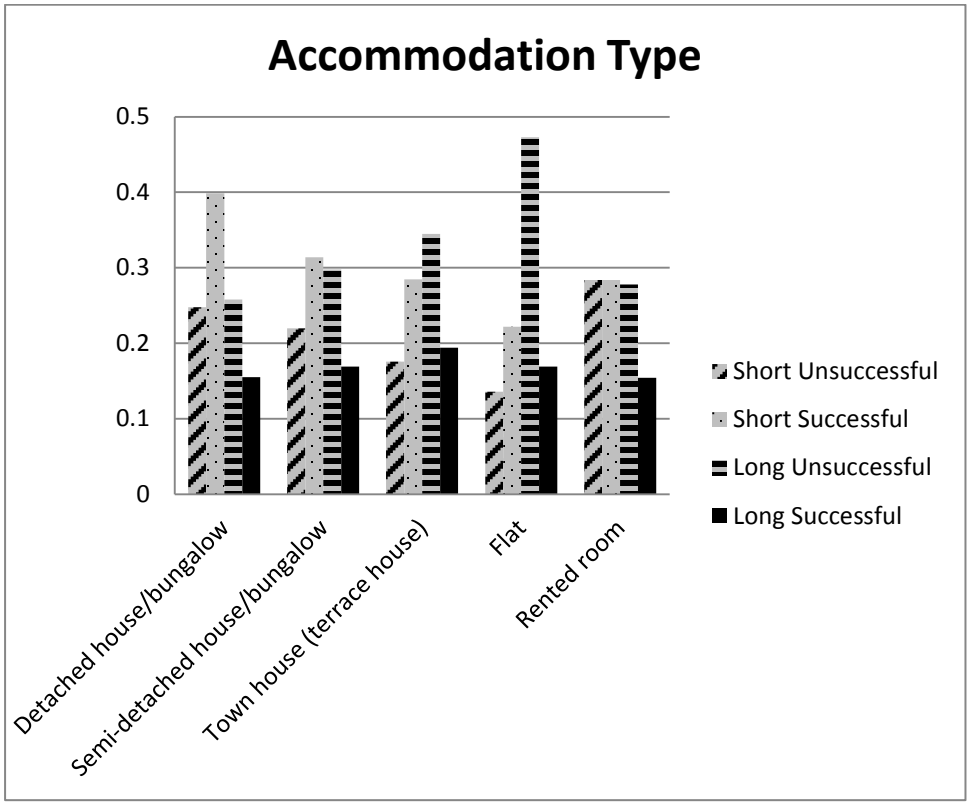
***p<0.001; **p<0.01; *p<0.05; ref – reference category.

Reference category for the response variable is Long Unsuccessful.

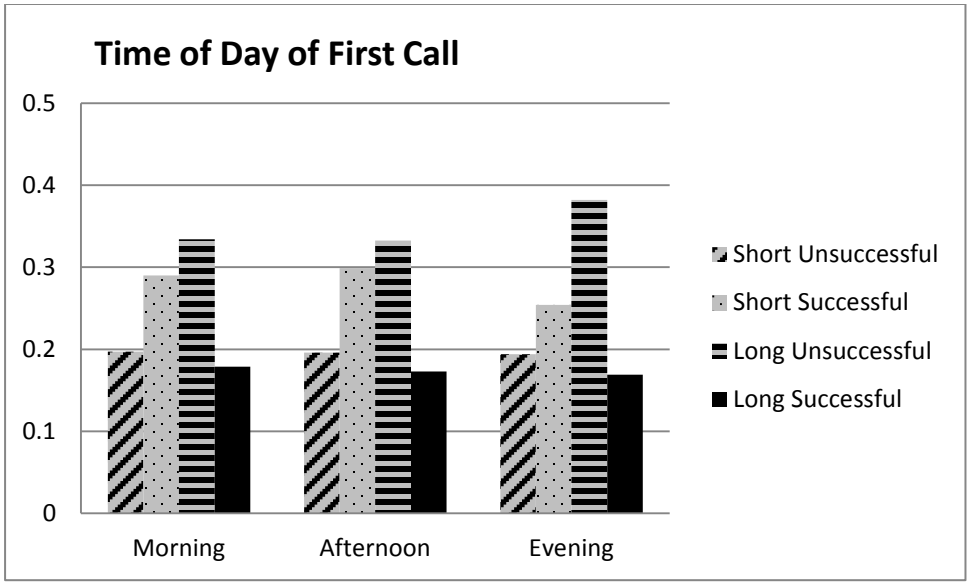
† The full model includes a constant and controls for geographic and design variables and all significant interviewer observation and call record variables.

Figure 2: Predicted probabilities for the final combined model (final outcome and length) for selected explanatory variables (Model 6) (The online appendix provides the figures in color.)

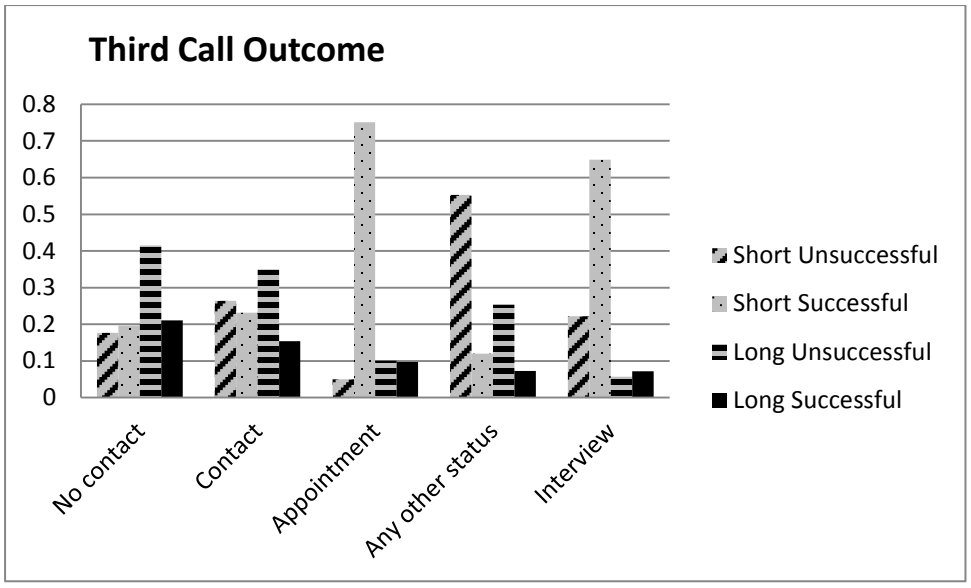
a.)



b.)



c.)



5. Conclusions and implications for survey practice

This paper presents an example of how to use paradata for interviewer-administered surveys, a topic often discussed in the literature. The paper employs a range of interviewer observation and call record variables to predict both final outcome and length of a call sequence early on in the data collection process. Models are developed prior to data collection and after the first, second and third call respectively to see if predictions of the models improve once more and more call record data are available. The models are developed both separately and jointly for the two outcomes of interest. The research was motivated by a recent sequence analysis (Durrant et al. 2014) which identified a categorisation of sequences based on length and outcome. Survey researchers have focussed on predicting outcome, either final outcome or outcome at each call (Groves and Couper 1998; Durrant, D'Arrigo, and Steele 2013). Here, this is extended to also include length of call sequence, a variable that so far has only been considered as an explanatory variable in non-response models.

The key findings from this analysis are:

1. Modelling outcome and length of call sequence jointly improves the fit of the model in comparison to the two separate models based on the pseudo-R^2 values and the classification tables. A comprehensive sensitivity analysis confirms that the main conclusions remain the same independent of the particular model specification.
2. The models proposed have the ability to predict the outcomes of interest reasonably well (length, final call outcome and the combined model) with a pseudo-R^2 of around 26% for the binary cases and 36% for the combined outcome. This is very high in a social science context and compares to values of around 3-8% for non-response models in the literature (Olson, Smyth, and Wood 2012; Olson and Groves 2012; West and Groves 2013). Classification tables show about 70% correctly classified cases for the binary cases (expected baseline is 50%) and even 52% for the multinomial case (expected baseline is 25%). Also, the AUC values for the ROC curves are around 0.75, which is higher than for other reported nonresponse models (Plewis et al, 2012) and significantly better than chance.
3. A number of variables are significant for both the model for length and the model for outcome. However, their effect may have opposite signs. For example, a household unlikely to have children has a higher probability of a shorter call sequence but a lower probability of response; a longer time in between calls may be associated with a shorter call sequence but with a lower probability of response.
4. *Call record variables:* The outcomes of previous calls, in particular the most recent call, are highly significant for both final response outcome and sequence length and their inclusion greatly improves the predictive power, with the R^2 increasing from less than 9% to above 26%. As one might expect, the prediction improves significantly when more and more call

outcomes are available. We found evidence that it is the most recent call outcome that may have the biggest influence, rather than the first call, for example. We found some indication that entering the raw outcome variables from each call (i.e. outcome of call 1, 2 and 3) as opposed to entering a summary measure, such as total number of non-contacts, improves the fit slightly. Often discussed in the literature as important variables, controlling for timing of calls has less impact on the performance of the model (R^2), although some of these variables are significant predictors throughout. Time between calls was often highly significant. Time of day was either not significant or only marginally significant. Interestingly, the day of the week was not found to be significant in the majority of models.

5. *Interviewer observation variables:* A number of interviewer observation variables (such as floor, car/van, child, relative condition of property) are found to be significant in all models. However, their contribution to the improvement of the model prediction is limited having controlled for previous call outcome(s). Adding interviewer observation variables to a model with, for example, only basic geographical information doubles or even quadruples the R^2 value. An indicator for the availability of interviewer observation variables may be used in practice to predict length and outcome.
6. *Basic geographic information*, available prior to the first call, are found to be significant variables for most models but are not found very effective in predicting final outcome and length of call sequence.

The findings highlight benefits and implications for survey practice, especially for adaptive and responsive survey designs, monitoring continually the streams of process data to alter the survey design during the course of data collection (Groves and Heeringa 2006). The paper

proposes a methodology that can be used and adapted by survey managers of other datasets, for both cross-sectional and longitudinal surveys. Different specifications of the dependent variables (e.g. Poisson, multinomial, ordinal, logistic) may be chosen depending on the survey design and the specific research questions. A sensitivity analysis is also recommended. We strongly encourage researchers to use sensitivity and the positive predicted value, classification tables and ROC curves rather than just the R^2 statistic to assess non-response models. The predictions from the two separate or the combined model provide *one* way of informing survey managers which households to follow up or where to stop calling. This paper focusses on identifying potentially long unsuccessful call sequences as early as possible during the data collection period. This guidance should, in practice, be supplemented with information about the impact of cases on nonresponse bias (and not just nonresponse rate) during data collection and further work in this area is currently being undertaken. Survey managers may wish to weigh up between the probability of a successful outcome versus sequence length. Ideally survey agencies may want to follow up those with potentially relatively short sequences and successful outcomes. Cases that are predicted to have long call sequences with many interviewer visits to a household may also be contacted by other means first (e.g. telephone) to avoid successive non-contacts at the doorstep. Often call record and interviewer observation variables can be supplemented further by additional linked data sources, such as from administrative or census data. Here, we only control for linked basic geographic information. Particular benefits of this work exist for data from longitudinal surveys since here the models can be, in principle, fitted for prediction in subsequent waves. Models may then be extended to include call record data and survey information from previous waves. Both of these further work strands are currently being explored.

The analysis indirectly assumes that all calls carry the same costs, e.g. staff and time resources and financial cost of a call. However, in practice some calls may be relatively inexpensive and may imply little extra burden, if, for example, a call can be made on the way to another household. The analysis could be extended to include such cost considerations. However, given the data we have available to us, we are unable to carry this out. The current work assumes a cut-off value π_0 to identify a household as a respondent or non-respondent based on their predicted probabilities. In practice, different values may be explored to optimise sensitivity and specificity in the specific survey setting considered. Other definitions of the dependent variables may also be considered (e.g. complete household response rather than 'at least one interview'). Due to the observational nature of the data we cannot establish causal links but merely associations between the response and the explanatory variables. However, this restriction is not a limitation as the main interest of the analysis is not to establish causal links but to compare performance of different models and to identify significant explanatory variables.

Supplementary Materials

Supplementary materials for this article may be found online at http://www.oxfordjournals.org/our_journals/jssam. The supplementary material includes exact wording of all variables used in the analysis, the distribution of explanatory variables broken down by the categories of the final dependent variables, the complete list of estimated coefficients for the two logistic regression models for length and (final) outcome and for the multinomial model and Figure 2 in color.

References

Agresti, A. (2013), *Categorical Data Analysis*, (3rd edition), New Jersey: John Wiley & Sons.

Altman, D. G. (1991), *Practical Statistics for Medical Research*, London: Champan & Hall.

Bates, N., Dahlhamer, J., and Singer, E. (2008), "Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse," *Journal of Official Statistics*, 24, 591–612.

Buck, N and McFall, S (2012), "Understanding Society: Design Overview", *Longitudinal and Life Course Studies*, 3, 1, 5-17.

Cameron, C.A. and Trivedi, P.K. (2013), Regression Analysis of Count Data, 2nd edition, New York, Cambridge Press.

Durrant, G. B., D'Arrigo, J., and Müller, G. (2013), "Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys," in *Improving Surveys with Paradata: Analytic Use of Process Information*, ed. Kreuter, F., New Jersey: John Wiley and Sons, pp. 281-308.

Durrant, G.B., D'Arrigo, J., and Steele, F. (2011), "Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences," *Journal of the Royal Statistical Society, Series A*, 174, 4, 1029-1049.

Durrant, G. B., D'Arrigo, J., and Steele, F. (2013), "Analysing Interviewer Call Record Data by Using a Multilevel Discrete-Time Event History Modelling Approach," *Journal of the Royal Statistical Society, Series A, Special issue: The Use of Paradata in Social Survey Research*, 176, 1, 251-269.

Durrant, G. B., Maslovskaya, O., and Smith, P. W. F. (2014), "Sequence Analysis as a Tool for Investigating Call Record Data," Working paper, University of Southampton.

Field, A. (2009), *Discovering statistics using SPSS* (3rd edition), Los Angeles: SAGE.

Groves, R. M., and Couper, M. P. (1996), "Contact-level Influences on Cooperation in Face-to-face Surveys," *Journal of Official Statistics*, 12, 63–83.

Groves, R. M., and Couper, M. P. (1998), *Nonresponse in Household Interview Surveys*, New York: John Wiley and Sons.

Groves, R. M., and Heeringa, S. G. (2006), "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society, Series A*, 169, 439–457.

Hanly, M. (2014), "Improving Nonresponse Bias Adjustments with Call Record Data," Paper to be presented at 25th International Workshop on Household Survey Nonresponse, Iceland.

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, pp. 221–233.

Kreuter, F., and Kohler, U. (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, 2, 203-226.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010), "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys," *Journal of the Royal Statistical Society, Series A*, 173, 389–407.

Krzanowski, W.J. and Hand, D.J. (2009), "*ROC Curves for Continuous Data*". Boca Raton, Florida: Chapman and Hall.

Long, J. S., and Freese, J. (2006), *Regression Models for Categorical Dependent Variables Using Stata* (2nd edition), Texas: STATA Press.

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

Olson, K., and Groves, R. M. (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., Smyth, J. D., and Wood, H. M. (2012), "Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination," *Public Opinion Quarterly*, 76, 611 – 635.

Pepe, M.S (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford: Oxford University Press.

Plewis, I., Ketende, S., and Calderwood, L. (2012), "Assessing the Accuracy of Response Propensities in Longitudinal Studies," *Survey Methodology*, 38, 2, 167-171.

Potthoff, R. F., Manton, K. G., and Woodbury, M. A. (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.

Sinibaldi, J. (2014), *Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications*, PhD Thesis, München: Ludwig-Maximilian-Universität.

Sinibaldi, J., Durrant, G. B., and Kreuter, F. (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.

Sinibaldi, J., Trappmann, M., and Kreuter, F. (2014), "Which is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations?" *Public Opinion Quarterly* (forthcoming).

Wagner, J. (2013a), "Adaptive Contact Strategies in Telephone and Face-to-Face Surveys," *Survey Research Methods*, 7, 1, 45-55.

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*, ed. Kreuter, F., New Jersey: John Wiley and Sons, pp. 145-170.

West, B. T., and Groves, R. M. (2013), "A Propensity-Adjusted Interviewer Performance Indicator," *Public Opinion Quarterly*, 77, 352-74.

White, H. (1980), "A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830.

White, H. (1984), *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press.

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

Online Appendix

Table A1: Exact wording of all variables used in the analysis.

Variable	Question
Used in the final models	
Months	Month of sample issue
Low density area for ethnic minorities	Low density area for ethnic minorities
Government Office Region (GOR)	Government Office Region of the address
Urban/rural	Is the address located in urban or rural area?
Barrier 2	Are any of these physical barriers to entry present at the address? Locked gates
Accommodation	Address dwelling type code
Floor	How many floors are there at the address?
Car/van	Based on your observation, is it likely that this address has a car or van?
Child	Based on your observation, is it likely that this address contains one or more children aged under 10 including babies?
Unkempt garden	Does the address have an unkempt garden?
Relative conditions of the address to other residential properties	How is the external condition of the address relative to other residential properties in the area?
Time of day call 1	Time call 1 started
Time of day call 2	Time call 2 started
Time of day call 3	Time call 3 started
Call 1 outcome	Status of call 1
Call 2 outcome	Status of call 2
Call 3 outcome	Status of call 3
Time between call 1 and call 2	Number of days between calls 1 and 2 (derived)
Time between call 2 and call 3	Number of days between calls 2 and 3 (derived)
Used in modelling but not significant	
Barrier 1	Are any of these physical barriers to entry present at the address? Locked common entrance
Barrier 3	Are any of these physical barriers to entry present at the address? Security staff or gatekeeper

Barrier 4	Are any of these physical barriers to entry present at the address? Entry phone access
Vicinity 1	Are any of the following present or within sight or hearing of the address? Boarded houses, abandoned buildings, demolished houses or demolished buildings
Vicinity 2	Are any of the following present or within sight or hearing of the address? Trash, litter or junk in street/road
Vicinity 3	Are any of the following present or within sight or hearing of the address? Heavy traffic on street/road
Conditions of residential property	Which of these best describes the condition of residential properties in the area?

Considered but not used in the final models shown

Sum of non-contacts	Sum of non-contacts in sequences (derived)
Sum of contacts	Sum of contacts in sequences (derived)
Sum of appointments	Sum of appointments in sequences (derived)
Sum of any other statuses	Sum of any other statuses in sequences (derived)
Time of day call 4	Time call 4 started
Day of week of call 1	Day of week call 1 was conducted (Day of week has 7 categories: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday)
Day of week of call 2	Day of week call 2 was conducted
Day of week of call 3	Day of week call 3 was conducted
Day of week of call 4	Day of week call 4 was conducted
Call 4 outcome	Status of call 4
Broken appointment between calls 1 and 2	Broken appointment between calls 1 and 2
Broken appointment between calls 2 and 3	Broken appointment between calls 2 and 3
Broken appointment between calls 3 and 4	Broken appointment between calls 3 and 4

Table A2: Distribution of explanatory variables broken down by the categories of the final dependent variables (a.) for (binary) length and final outcome and b.) for the combined dependent variable of length and outcome).

a.)

Variables	Length		(Final) Outcome	
	Short (up to 6 calls)	**Long (more than 6 calls)**	**No interviews after call 3**	**At least one interview after call 3**
Geographic and design				
Months				
January-June year 1	3,209 (51.3%)	3,044 (48.7%)	3,253 (52.0%)	3,000 (48.0%)
July-December year 1	2,897 (49.6%)	2,946 (50.4%)	2,978 (51.0%)	2,865 (49.0%)
January-June year 2	3,095 (47.9%)	3,371 (52.1%)	3,537 (54.7%)	2,929 (45.3%)
July-December year 2	3,152 (46.4%)	3,644 (53.6%)	3,797 (55.9%)	2,999 (44.1%)
Low density area for ethnic minorities				
No	4,513 (42.9%)	6,000 (57.1%)	5,927 (56.4%)	4,586 (43.6%)
Yes	7,840 (52.8%)	7,005 (47.2%)	7,638 (51.5%)	7,207 (48.5%)
Government Office Region (GOR)				
North East	540 (47.0%)	609 (53.0%)	551 (48.0%)	598 (52.0%)
North West	1,405 (50.2%)	1,395 (49.8%)	1,404 (50.1%)	1,396 (49.9%)
Yorkshire and the Humber	1,119 (53.1%)	987 (46.9%)	1,084 (51.5%)	1,022 (48.5%)
East Midlands	952 (52.8%)	850 (47.2%)	836 (46.4%)	966 (53.6%)
West Midlands	1,195 (51.4%)	1,129 (48.6%)	1,253 (53.9%)	1,071 (46.1%)
East of England	1,254 (50.3%)	1,241 (49.7%)	1,326 (53.1%)	1,169 (46.9%)
London	1,279 (34.7%)	2,402 (65.3%)	2,312 (62.8%)	1,369 (37.2%)
South East	1,839 (51.0%)	1,770 (49.0%)	1,943 (53.8%)	1,666 (46.2%)
South West	1,062 (53.6%)	919 (46.4%)	1,025 (51.7%)	956 (48.3%)
Wales	594 (52.4%)	540 (47.6%)	574 (50.6%)	560 (49.4%)
Scotland	1,114 (48.9%)	1,163 (51.1%)	1,257 (55.2%)	1,020 (44.8%)
Urban/rural				
Urban area	9,747 (46.9%)	11,045 (53.1%)	11,255 (54.1%)	9,537 (45.9%)
Rural area	2,606 (57.1%)	1,960 (42.9%)	2,310 (50.6%)	2,256 (49.4%)
Interviewer observations				
Accommodation				
Detached	3,016 (58.7%)	2,124 (41.3%)	2,598 (50.5%)	2,542 (49.5%)

house/bungalow				
Semi-detached house/bungalow	3,921 (53.3%)	3,430 (46.7%)	3,803 (51.7%)	3,548 (48.3%)
Town house (Terrace /end terrace)	3,483 (46.1%)	4,080 (53.9%)	3,943 (52.1%)	3,620 (47.9%)
Flat (incl. maisonette) purpose built or converted	1,841 (35.8%)	3,301 (64.2%)	3,130 (60.9%)	2,012 (39.1%)
Rented rooms (bedsitters), dwellings with business and sheltered accommodation	92 (56.8%)	70 (43.2%)	91 (56.2%)	71 (43.8%)
Floor				
0 floors	76 (67.9%)	36 (32.1%)	49 (43.8%)	63 (56.2%)
1 floor	1,824 (52.5%)	1,651 (47.5%)	1,857 (53.4%)	1,618 (46.6%)
2 floors	9,186 (50.7%)	8,944 (49.3%)	9,415 (51.9%)	8,715 (48.1%)
3 floors	814 (38.0%)	1,328 (62.0%)	1,205 (56.3%)	937 (43.7%)
4 floors and above	453 (30.2%)	1,046 (69.8%)	1,039 (69.3%)	460 (30.7%)
Car/van				
Definitely has a car/van	5,076 (55.1%)	4,132 (44.9%)	4,337 (47.1%)	4,871 (52.9%)
Likely	2,636 (50.2%)	2,610 (49.8%)	2,790 (53.2%)	2,456 (46.8%)
Unlikely	527 (50.5%)	516 (49.5%)	647 (62.0%)	396 (38.0%)
Definitely does not have a car/van	417 (62.2%)	253 (37.8%)	206 (30.7%)	464 (69.3%)
Cannot tell from observation	3,697 (40.2%)	5,494 (59.8%)	5,585 (60.8%)	3,606 (39.2%)
Child				
Definitely has a child/children aged under 10	853 (51.3%)	811 (48.7%)	693 (41.6%)	971 (58.4%)
Likely	1,128 (51.2%)	1,073 (48.8%)	1,008 (45.8%)	1,193 (54.2%)
Unlikely	2,630 (51.0%)	2,522 (49.0%)	2,866 (55.6%)	2,286 (44.4%)
Definitely does not have a child/children aged under 10	2,133 (54.9%)	1,755 (45.1%)	1,864 (47.9%)	2,024 (52.1%)
Cannot tell from observation	5,609 (45.0%)	6,844 (55.0%)	7,134 (57.3%)	5,319 (42.7%)
Unkempt garden				
Yes	1,001 (42.9%)	1,335 (57.1%)	1,249 (53.5%)	1,087 (46.5%)
No	9,007 (52.6%)	8,121 (47.4%)	8,891 (51.9%)	8,237 (48.1%)
No obvious garden	2,345 (39.8%)	3,549 (60.2%)	3,425 (58.1%)	2,469 (41.9%)

Relative conditions of the address to other residential properties				
Better	1,054 (53.5%)	917 (46.5%)	912 (46.3%)	1,059 (53.7%)
About the same	10,517 (48.9%)	10,981 (51.5%)	11,529 (53.6%)	9,969 (46.4%)
Worse	743 (41.4%)	1,050 (58.6%)	1,047 (58.4%)	746 (41.6%)
Unable to obtain information	39 (40.6%)	57 (59.4%)	77 (80.2%)	19 (19.8%)
Call Record Variables				
Time of day call 1				
Morning (0-12.00)	2,280 (48.7%)	2,402 (51.3%)	2,485 (53.1%)	2,197 (46.9%)
Afternoon (12.00-17.00)	8,485 (49.5%)	8,647 (50.5%)	9,037 (52.7%)	8,095 (47.3%)
Evening (17.00-24.00)	1,588 (44.8%)	1,956 (55.2%)	2,043 (57.6%)	1,501 (42.4%)
Time of day call 2				
Morning (0-12.00)	2,377 (50.7%)	2,314 (49.3%)	2,524 (53.8%)	2,167 (46.2%)
Afternoon (12.00-17.00)	6,653 (48.7%)	7,016 (51.3%)	7,198 (52.7%)	6,471 (47.3%)
Evening (17.00-24.00)	3,323 (47.5%)	3,675 (52.5%)	3,843 (54.9%)	3,155 (45.1%)
Time between call 1 and call 2				
Time of day call 3				
Morning(0-12.00)	2,346 (50.4%)	2,313 (49.6%)	2,515 (54.0%)	2,144 (46.0%)
Afternoon (12.00-17.00)	5,801 (49.3%)	5,974 (50.7%)	6,325 (53.7%)	5,450 (46.3%)
Evening (17.00-24.00)	4,206 (47.1%)	4,718 (52.9%)	4,725 (52.9%)	4,199 (47.1%)
Time between call 2 and call 3				
Call 1 outcome				
No contact	7,876 (45.0%)	9,640 (55.0%)	9,135 (52.2%)	8,381 (47.8%)
Contact made	2,957 (55.4%)	2,385 (44.6%)	3,136 (58.7%)	2,206 (41.3%)
Appointment made	909 (58.1%)	655 (41.9%)	680 (43.5%)	884 (56.5%)
Any other status	521 (64.2%)	291 (35.8%)	564 (69.5%)	248 (30.5%)
Interview done	90 (72.6%)	34 (27.4%)	50 (40.3%)	74 (59.7%)
Call 2 outcome				
No contact	7,225 (43.2%)	9,513 (56.8%)	8,900 (53.2%)	7,838 (46.8%)
Contact made	2,695 (53.2%)	2,372 (46.8%)	2,909 (57.4%)	2,158 (42.6%)
Appointment	1,215 (62.7%)	722 (37.3%)	750 (38.7%)	1187 (61.3%)

made				
Any other status	808 (71.8%)	317 (28.2%)	843 (74.9%)	282 (25.1%)
Interview done	410 (83.5%)	81 (16.5%)	163 (33.2%)	328 (66.8%)
Call 3 outcome				
No contact	5,706 (37.4%)	9,547 (62.6%)	9,034 (59.2%)	6,219 (40.8%)
Contact made	2,196 (49.6%)	2,230 (50.4%)	2,717 (61.4%)	1,709 (38.6%)
Appointment made or interview or complete	2,793 (80.1%)	692 (19.9%)	528 (15.2%)	2,957 (84.8%)
Any other status	860 (67.3%)	418 (32.7%)	1,031 (80.7%)	247 (19.3%)
Interview done	798 (87.1%)	118 (12.9%)	255 7.8%)	661 (72.2%)

b.)

Variables	Short Unsuccessful	Short Successful	Long Unsuccessful	Long Successful
Geographic and design				
Months				
January-June year 1	1,279 (20.5%)	1,930 (30.9%)	1,974 (31.6%)	1,070 (17.1%)
July-December year 1	1,061 (18.2%)	1,836 (31.4%)	1,917 (32.8%)	1,029 (17.6%)
January-June year 2	1,271 (19.7%)	1,824 (28.2%)	2,266 (35.0%)	1,105 (17.1%)
July-December year 2	1,351 (19.9%)	1,801 (26.5%)	2,446 (36.0%)	1,198 (17.6%)
Low density area for ethnic minorities				
No	1,821 (17.3%)	2,692 (25.6%)	4,106 (39.1%)	1,894 (18.0%)
Yes	3,141 (21.2%)	4,699 (31.7%)	4,497 (30.3%)	2,508 (16.9%)
Government Office Region (GOR)				
North East	192 (16.7%)	348 (30.3%)	359 (31.2%)	250 (21.8%)
North West	514 (18.4%)	891 (31.8%)	890 (31.8%)	505 (18.0%)
Yorkshire and the Humber	458 (21.7%)	661 (31.4%)	626 (29.7%)	361 (17.1%)
East Midlands	342 (19.0%)	610 (33.9%)	494 (27.4%)	356 (19.8%)
West Midlands	502 (21.6%)	693 (29.8%)	751 (32.3%)	378 (16.3%)
East of England	481 (19.3%)	773 (31.0%)	845 (33.9%)	396 (15.9%)
London	557 (15.1%)	722 (19.6%)	1,755 (47.7%)	647 (17.6%)
South East	792 (21.9%)	1,047 (29.0%)	1,151 (31.9%)	619 (17.2%)
South West	407 (20.5%)	655 (33.1%)	618 (31.2%)	301 (15.2%)
Wales	243 (21.4%)	351 (31.0%)	331 (29.2%)	209 (18.4%)
Scotland	474 (20.8%)	640 (28.1%)	783 (34.4%)	380 (16.7%)
Urban/rural				
Urban area	3,886 (18.7%)	5,861 (28.2%)	7,369 (35.4%)	3,676 (17.7%)

Rural area	1,076 (23.6%)	1,530 (33.5%)	1,234 (27.0%)	726 (15.9%)
Interviewer observations				
Barrier 2 (locked gates)				
Not mentioned	4,882 (19.6%)	7,311 (29.3%)	8,395 (33.7%)	4,324 (17.4%)
Mentioned	80 (17.9%)	80 (17.9%)	208 (46.6%)	78 (17.5%)
Accommodation				
Detached house/bungalow	1,273 (24.8%)	1,743 (33.9%)	1,325 (25.8%)	799 (15.5%)
Semi-detached house/bungalow	1,615 (22.0%)	2,306 (31.4%)	2,188 (29.8%)	1,242 (16.9%)
Town house (Terrace /end terrace)	1,330 (17.6%)	2,153 (28.5%)	2,613 (34.5%)	1,467 (19.4%)
Flats (incl. maisonettes) purpose built or converted	698 (13.6%)	1,143 (22.2%)	2,432 (47.3%)	869 (16.9%)
Rented rooms (bedsitters), dwellings with business and sheltered accommodation	46 (28.4%)	46 (28.4%)	45 (27.8%)	25 (15.4%)
Floor				
0 floors	32 (38.6%)	44 (39.3%)	17 (15.2%)	19 (17.0%)
1 floor	771 (22.2%)	1,053 (30.3%)	1,086 (31.3%)	565 (16.3%)
2 floors	3,666 (20.2%)	5,520 (30.4%)	5,749 (31.7%)	3,195 (17.6%)
3 floors	288 (13.4%)	526 (24.6%)	917 (42.8%)	411 (19.2%)
4 floors and above	205 (13.7%)	248 (16.5%)	834 (55.6%)	212 (14.1%)
Car/van				
Definitely has a car/van	1,855 (20.1%)	3,221 (35.0%)	2,482 (27.0%)	1,650 (17.9%)
Likely	1,092 (20.8%)	1,544 (29.4%)	1,698 (32.4%)	912 (17.4%)
Unlikely	269 (25.8%)	258 (24.7%)	378 (36.2%)	138 (13.2%)
Definitely does not have a car/van	103 (15.4%)	314 (46.9%)	103 (15.4%)	150 (22.4%)
Cannot tell from observation	1,643 (17.9%)	2,054 (22.3%)	3,942 (42.9%)	1,552 (16.9%)
Child				
Definitely has a child/children aged under 10	247 (14.8%)	606 (36.4%)	446 (26.8%)	365 (21.9%)
Likely	360 (16.4%)	768 (34.9%)	648 (29.4%)	425 (19.3%)
Unlikely	1,172 (22.7%)	1,458 (28.3%)	1,694 (32.9%)	828 (16.1%)
Definitely does not have a child/children aged under 10	823 (21.2%)	1,310 (33.7%)	1,041 (26.8%)	714 (18.4%)
Cannot tell from	2,360 (19.0%)	3,249 (26.1%)	4,774 (38.3%)	2,070 (16.6%)

observation				
Unkempt garden				
Yes	359 (15.4%)	642 (27.5%)	890 (38.1%)	445 (19.0%)
No	3,670 (21.4%)	5,337 (31.2%)	5,221 (30.5%)	2,900 (16.9%)
No obvious garden	933 (15.8%)	1,412 (24.0%)	2,492 (42.3%)	1,057 (17.9%)
Relative conditions of the address to other residential properties				
Better	351 (17.8%)	703 (35.7%)	561 (28.5%)	356 (18.1%)
About the same	4,272 (19.9%)	6,245 (29.0%)	7,257 (33.8%)	3,724 (17.3%)
Worse	314 (17.5%)	429 (23.9%)	733 (40.9%)	317 (17.7%)
Unable to obtain information	25 (26.0%)	14 (14.6%)	52 (54.2%)	5 (5.2%)
Call Record Variables				
Time of day call 1				
Morning (0-12.00)	921 (19.7%)	1,359 (29.0%)	1,564 (33.4%)	838 (17.9%)
Afternoon (12.00-17.00)	3,353 (19.6%)	5,132 (30.0%)	5,684 (33.2%)	2,963 (17.3%)
Evening (17.00-24.00)	688 (19.4%)	900 (25.4%)	1,355 (38.2%)	601 (17.0%)
Time of day call 2				
Morning (0-12.00)	961 (20.5%)	1,416 (30.2%)	1,563 (33.3%)	751 (16.0%)
Afternoon (12.00-17.00)	2,682 (19.6%)	3,971 (29.1%)	4,516 (33.0%)	2,500 (18.3%)
Evening (17.00-24.00)	1,319 (18.8%)	2,004 (28.6%)	2,524 (36.1%)	1,151 (16.4%)
Time between call 1 and call 2				
Time of day call 3				
Morning(0-12.00)	979 (21.0%)	1,367 (29.3%)	1,536 (33.0%)	777 (16.7%)
Afternoon (12.00-17.00)	2,455 (20.8%)	3,346 (28.4%)	3,870 (32.9%)	2,104 (17.9%)
Evening (17.00-24.00)	1,528 (17.1%)	2,678 (30.0%)	3,197 (35.8%)	1,521 (17.0%)
Time between call 2 and call 3				
Call 1 outcome				
No contact	2,735 (15.6%)	5,141 (29.4%)	6,400 (36.5%)	3,240 (18.5%)
Contact made	1,527 (28.6%)	1,430 (26.8%)	1,609 (30.1%)	776 (14.5%)
Appointment made	310 (19.8%)	599 (38.3%)	370 (23.7%)	285 (18.2%)
Any other status	355 (43.7%)	166 (20.4%)	209 (25.7%)	82 (10.1%)

Interview done	35 (28.2%)	55 (44.4%)	15 (12.1%)	19 (15.3%)
Call 2 outcome				
No contact	2,603 (15.6%)	4,622 (27.6%)	6,297 (37.6%)	3,216 (19.2%)
Contact made	1,283 (25.3%)	1,412 (27.9%)	1,626 (32.1%)	746 (14.7%)
Appointment made	332 (17.1%)	883 (45.6%)	418 (21.6%)	304 (15.7%)
Any other status	611 (54.3%)	197 (17.5%)	232 (20.6%)	85 (7.6%)
Interview done	133 (27.1%)	277 (56.4%)	30 (6.1%)	51 (10.4%)
Call 3 outcome				
No contact	2,706 (17.7%)	3,000 (19.7%)	6,328 (41.5%)	3,219 (21.1%)
Contact made	1,170 (26.4%)	1,026 (23.2%)	1,547 (35.0%)	683 (15.4%)
Appointment made	176 (5.1%)	2,617 (75.1%)	352 (10.1%)	340 (9.8%)
Any other status	707 (55.3%)	153 (12.0%)	324 (25.4%)	94 (7.4%)
Interview done	203 (22.2%)	595 (65.0%)	52 (5.7%)	66 (7.2%)

Table A3:

Estimated coefficients for the two logistic regression models for length and (final) outcome including geographic and design variables, interviewer observation variables and call record variables comprising timing and outcome of the call(s) (Model 8).

	Model for Length			Model for (final) outcome		
Variable	**B**	**Robust SE**	**OR**	**β**	**Robust SE**	**OR**
Constant	-1.087	0.149		0.363	0.149	
Geographic and design variables						
Months						
January-June year 1 (ref)	0.000		1.000	0.000		1.000
July-December year 1	-0.099	0.040	0.905*	0.154	0.040	1.167***
January-June year 2	-0.213	0.039	0.808***	-0.001	0.039	0.999
July-December year 2	-0.319	0.039	0.726***	0.016	0.039	1.017
Low density area for ethnic minorities						
No (ref)	0.000		1.000			
Yes	0.086	0.036	1.090*			
Government Office Region (GOR)						
North East (ref)	0.000		1.000	0.000		1.000
North West	0.091	0.079	1.095	-0.164	0.077	0.848*
Yorkshire and the Humber	0.136	0.082	1.146	-0.168	0.080	0.845*
East Midlands	0.048	0.085	1.050	0.072	0.083	1.075
West Midlands	0.084	0.082	1.088	-0.244	0.079	0.783**
East of England	0.016	0.077	1.016	-0.212	0.078	0.809**
London	-0.282	0.082	0.754**	-0.449	0.075	0.638***
South East	0.061	0.077	1.063	-0.250	0.074	0.779**
South West	0.237	0.083	1.267**	-0.184	0.081	0.832*
Wales	0.156	0.093	1.169	-0.151	0.092	0.859
Scotland	0.182	0.083	1.199*	-0.334	0.080	0.716***
Urban/rural						
Urban area (ref)	0.000		1.000			
Rural area	0.121	0.040	1.128**			
Interviewer observations variables						
Accommodation						
Detached house/bungalow (ref)	0.000		1.000			
Semi-detached house/bungalow	-0.062	0.042	0.940			
Town house (terrace)	-0.207	0.046	0.813***			
Flat (incl. maisonette) purpose built or converted	-0.296	0.061	0.744***			
Rented room (bedsitter), dwellings with business and sheltered accommodation	0.257	0.184	1.293			

Floor						
0 floors	0.809	0.236	2.247**	0.806	0.213	2.238***
1 floor	0.318	0.078	1.374***	0.430	0.073	1.538***
2 floors	0.196	0.075	1.216**	0.402	0.065	1.495***
3 floors	0.051	0.079	1.052	0.413	0.077	1.511***
4 floors and above (ref)	0.000		1.000	0.000		1.000
Car/van						
Definitely has a car/van (ref)	0.000		1.000	0.000		1.000
Likely	-0.039	0.041	0.962	-0.189	0.039	0.827***
Unlikely	0.153	0.078	1.165	-0.452	0.075	0.636***
Definitely does not have a car/van	0.666	0.095	1.947***	0.654	0.095	1.924***
Cannot tell from observation	-0.137	0.041	0.872**	-0.400	0.036	0.670***
Child						
Definitely has a child/children aged under 10 (ref)	0.000		1.000	0.000		1.000
Likely	0.013	0.075	1.013	-0.099	0.075	0.906
Unlikely	0.177	0.068	1.193**	-0.272	0.065	0.762***
Definitely does not have a child/children aged under 10	0.209	0.068	1.232**	-0.163	0.066	0.849*
Cannot tell from observation	0.058	0.062	1.060	-0.312	0.061	0.732***
Unkempt garden						
Yes (ref)	0.000		1.000			
No	0.203	0.056	1.225***			
No obvious garden	0.113	0.062	1.120			
Relative conditions of the address to other residential properties						
Better (ref)	0.000		1.000	0.000		1.000
About the same	-0.047	0.054	0.954	-0.242	0.053	0.785***
Worse	-0.202	0.078	0.903*	-0.440	0.074	0.644***
Unable to obtain information	0.008	0.239	1.008	-1.248	0.260	0.287***
Call Record Variables						
Time of day call 1						
Morning (0.00-12.00) (ref)				0.000		1.000
Afternoon (12.00-17.00)				0.010	0.036	1.011
Evening (17.00-24.00)				-0.130	0.050	0.878**
Time of day call 2						
Morning (0.00-12.00) (ref)	0.000		1.000			
Afternoon (12.00-17.00)	-0.064	0.038	0.937			
Evening (17.00-24.00)	-0.102	0.043	0.903*			
Time of day call 3						
Morning (0.00-12.00) (ref)						
Afternoon (12.00-17.00)	-0.015	0.039	0.985			
Evening (17.00-24.00)	-0.094	0.041	0.912*			
Time between call 1 and call 2	0.026	0.002	1.026***	-0.026	0.002	0.974***

Time between call 2 and call 3		0.030	0.001	1.030***	-0.028	0.002	0.973***
Call 1 outcome							
	No contact (ref)	0.000		1.000	0.000		1.000
	Contact made	0.081	0.037	1.085*	-0.098	0.037	0.907**
	Appointment made	0.025	0.069	1.026	0.242	0.065	1.274***
	Any other status	0.124	0.095	1.133	-0.035	0.096	0.966
	Interview done	1.022	0.240	2.780***	0.372	0.207	1.450
Call 2 outcome							
	No contact (ref)	0.000		1.000	0.000		1.000
	Contact made	0.095	0.038	1.099*	-0.062	0.037	0.940
	Appointment made	0.205	0.063	1.228**	0.458	0.062	1.581***
	Any other status	0.293	0.088	1.340**	-0.024	0.092	0.977
	Interview done	1.539	0.143	4.662***	0.517	0.116	1.677***
Call 3 outcome							
	No contact (ref)	0.000		1.000	0.000		1.000
	Contact made	0.499	0.037	1.645***	-0.148	0.037	0.862***
	Appointment made	2.000	0.048	7.389***	2.024	0.053	7.568***
	Any other status	1.227	0.065	3.410***	-1.185	0.074	0.305***
	Interview done	2.352	0.110	10.511***	0.860	0.087	2,364***

***p<0.001; **p<0.01; *p<0.05; ref – reference category

Table A4

Estimated coefficients for the multinomial regression model for the combined dependent variable (length and final outcome) including geographic and design variables, interviewer observation variables and call record variables comprising timing and outcome of the call(s) (Model 8).

	Short Unsuccessful		Short Successful		Long Successful	
Variable	**β**	**Robust SE**	**β**	**Robust SE**	**β**	**Robust SE**
Constant	-1.832***	0.216	-0.396*	0.192	-0.018	0.198
Geographic and design variables						
Months						
January-June year 1 (ref)	0.000		0.000		0.000	
July-December year 1	-0.316***	0.058	0.060	0.051	0.032	0.055
January-June year 2	-0.354***	0.055	-0.157**	0.051	-0.071	0.054
July-December year 2	-0.424***	0.055	-0.237***	0.051	-0.001	0.053
Government Office Region (GOR)						
North East (ref)	0.000		0.000		0.000	
North West	0.020	0.112	-0.049	0.101	-0.259*	0.101
Yorkshire and the Humber	0.168	0.116	-0.029	0.105	-0.212	0.107
East Midlands	-0.040	0.122	0.105	0.109	0.047	0.110
West Midlands	0.065	0.114	-0.137	0.104	-0.305**	0.104
East of England	-0.130	0.114	-0.149	0.102	-0.358**	0.104
London	-0.323**	0.111	-0.591***	0.100	-0.439***	0.098
South East	0.123	0.108	-0.168	0.098	-0.245*	0.098
South West	0.195	0.118	0.046	0.106	-0.337**	0.110
Wales	0.271*	0.133	-0.007	0.120	-0.125	0.122
Scotland	0.389**	0.117	-0.143	0.107	-0.280**	0.107
Urban/rural						
Urban area (ref)	0.000		0.000		0.000	
Rural area	0.183**	0.054	0.136**	0.050	0.015	0.057
Interviewer observations variables						
Barrier 2 (locked gates)						
Not mentioned (ref)	0.000		0.000		0.000	
Mentioned	-0.025	0.151	-0.515***	0.147	-0.147	0.141
Accommodation						
Detached house/bungalow (ref)	0.000		0.000		0.000	
Semi-detached house/bungalow	-0.073	0.059	-0.072	0.054	-0.011	0.060
Town house (Terrace)	-0.320***	0.066	-0.101	0.059	0.089	0.064
Flat (incl.	-0.604***	0.086	-0.147	0.077	-0.111	0.082

maisonette) purpose built or converted						
Rented room (bedsitter), dwellings with business and sheltered accommodation	0.363	0.242	0.268	0.237	0.163	0.266
Floor						
0 floors	0.997**	0.342	1.343***	0.317	1.084**	0.350
1 floor	0.248*	0.111	0.524***	0.101	0.366***	0.104
2 floors	0.135	0.107	0.383***	0.097	0.349***	0.098
3 floors	-0.092	0.114	0.311**	0.102	0.394***	0.100
4 floors and above (ref)	0.000		0.000		0.000	
Car/van						
Definitely has a car/van (ref)	0.000		0.000		0.000	
Likely	0.067	0.058	-0.206***	0.052	-0.147**	0.056
Unlikely	0.329**	0.105	-0.271**	0.101	-0.500***	0.110
Definitely does not have a car/van	0.838***	0.157	1.101***	0.133	-0.746***	0.138
Cannot tell from observation	0.043	0.058	-0.468***	0.053	0.328***	0.055
Child						
Definitely has a child/children aged under 10 (ref)	0.000		0.000		0.000	
Likely	-0.056	0.115	-0.055	0.096	-0.162	0.098
Unlikely	0.237*	0.099	-0.068	0.084	-0.330***	0.086
Definitely does not have a child/children aged under 10	0.314**	0.102	0.029	0.086	-0.164	0.087
Cannot tell from observation	0.095	0.079	-0.185*	0.079	-0.370***	0.080
Unkempt garden						
Yes (ref)	0.000		0.000		0.000	
No	0.311***	0.077	0.200**	0.072	0.094	0.071
No obvious garden	0.212*	0.085	0.071	0.079	0.047	0.077
Relative conditions of the address to other residential properties						
Better (ref)	0.000		0.000		0.000	
About the same	0.093	0.079	-0.255***	0.067	-0.162*	0.073
Worse	0.014	0.110	-0.525***	0.101	-0.288**	0.105
Unable to obtain information	0.337	0.313	-0.872**	0.311	-1.540**	0.480
Call Record Variables						
Time of day call 1						
Morning (0.00-12.00) (ref)	0.000		0.000		0.000	

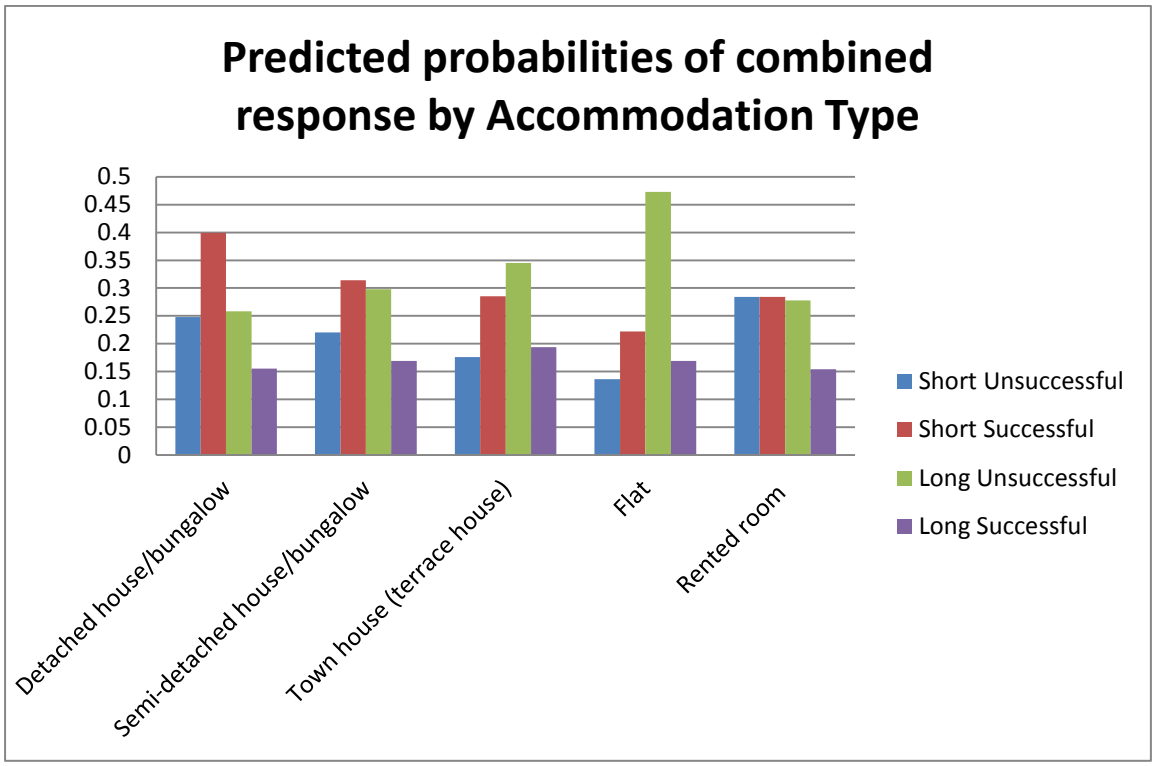
Afternoon (12.00-17.00)	0.018	0.052	0.034	0.047	-0.021	0.049
Evening (17.00-24.00)	-0.081	0.071	-0.180**	0.065	-0.119	0.067
Time of day call 2						
Morning (0.00-12.00) (ref)	0.000		0.000		0.000	
Afternoon (12.00-17.00)	-0.023	0.053	-0.012	0.048	0.141**	0.052
Evening (17.00-24.00)	-0.126*	0.059	-0.072	0.054	0.011	0.059
Time of day call 3						
Morning (0.00-12.00) (ref)	0.000		0.000		0.000	
Afternoon (12.00-17.00)	-0.025	0.054	0.028	0.050	0.058	0.053
Evening (17.00-24.00)	-0.127*	0.057	-0.083	0.052	-0.034	0.055
Time between call 1 and call 2	0.034***	0.002	-0.002	0.002	-0.024***	0.003
Time between call 2 and call 3	0.038***	0.002	-0.001	0.002	-0.028***	0.003
Call 1 outcome						
No contact (ref)	0.000		0.000		0.000	
Contact made	0.279***	0.050	-0.038	0.048	0.035	0.052
Appointment made	0.124	0.094	0.219*	0.088	0.337***	0.087
Any other status	0.292*	0.118	0.055	0.130	0.138	0.145
Interview done	1.510***	0.335	1.320***	0.327	0.821*	0.350
Call 2 outcome						
No contact (ref)	0.000		0.000		0.000	
Contact made	0.198***	0.052	0.012	0.049	-0.024	0.052
Appointment made	0.032	0.089	0.516***	0.081	0.316***	0.086
Any other status	0.436***	0.107	0.202	0.124	0.228	0.147
Interview done	2.056	0.221	2.043***	0.210	1.012***	0.238
Call 3 outcome						
No contact (ref)	0.000		0.000		0.000	
Contact made	0.618***	0.050	0.273***	0.049	-0.203***	0.052
Appointment made	0.187	0.102	2.683***	0.063	0.568***	0.080
Any other status	1.780***	0.077	-0.125	0.102	-0.661***	0.119
Interview done	2.344***	0.169	2.704***	0.157	0.556**	0.195

***p<0.001; **p<0.01; *p<0.05; ref – reference category.

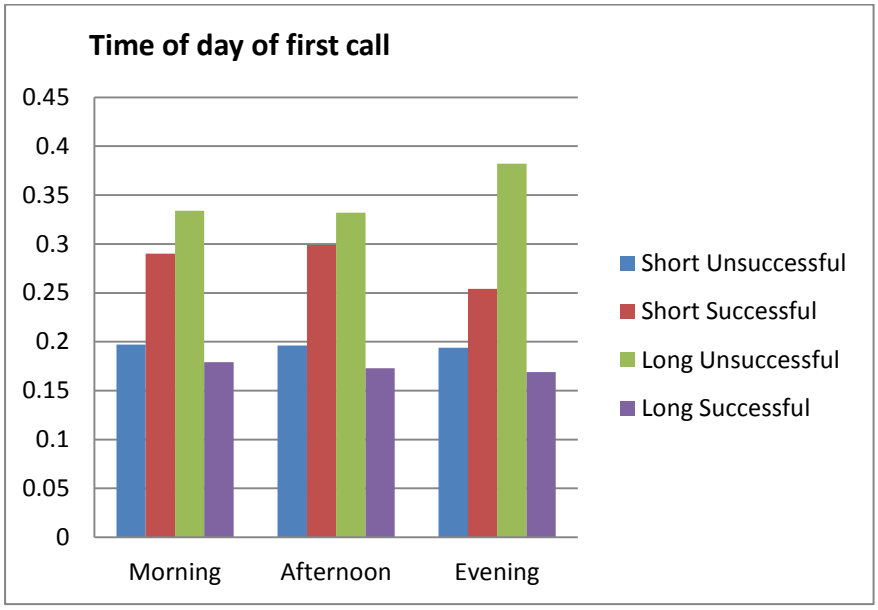
Reference category for the response variable is Long Unsuccessful.

Figure A1: Figure 2 from text in color.
Predicted probabilities for the final combined model (final outcome and length) for selected explanatory variables (one variable from each group of explanatory variables in Model 8)

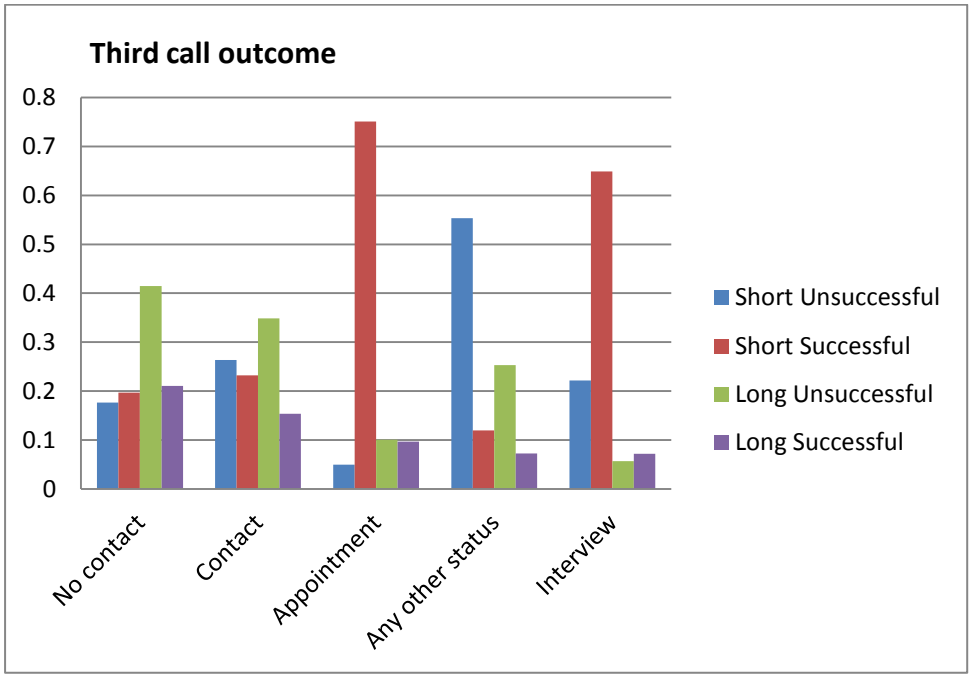
a.)



b.)



c.)


Third call outcome
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
No contact
Contact
Appointment
Any other status
Interview
Short Unsuccessful
Short Successful
Long Unsuccessful
Long Successful