Working paper – please do not cite without permission of authors

Assessing nonresponse bias using call record data with applications to a longitudinal study

Solange Correa, Gabriele B. Durrant and Peter W. F. Smith

University of Southampton, UK

Address for correspondence:

Solange Correa-Onel
Department of Social Statistics and Demography
University of Southampton
Southampton
SO17 1BJ, UK.

E-mail: s.correa-onel@soton.ac.uk

Summary. A method to monitor survey outcomes during fieldwork is proposed. The approach assesses nonresponse bias using call record data by comparing estimated and "true" distributions of specific survey variables at each call attempt using dissimilarity indices. These are compared with other survey quality indicators such as response rate, nonresponse bias, R-indicators, coefficients of variation, partial R-indicators and partial coefficients of variation. Empirical analyses are conducted using data from Understanding Society – the UK Household Longitudinal Study. Results show that survey estimates tend to stabilise after around 5 call attempts. The study demonstrates that a number of indicators commonly used, although adequate to assess nonresponse bias after data collection, may not be effective in capturing nonresponse bias during the call process. The study concludes that dissimilarity indices and coefficients of variation exhibit best properties. This research has implications for responsive and adaptive survey designs.

Keywords: **N**onresponse bias monitoring; paradata; survey quality indicator; Understanding Society

1. Introduction

In recent years the focus in survey research has shifted from simply monitoring response rates to also monitoring and assessing nonresponse bias of key survey estimates (Groves and Couper, 1998; Groves, 2006; Groves and Heeringa, 2006; Merkle and Edelman, 2009; Durrant *et al.*, 2013; Kreuter, 2013). To be able to do this, fully observed data for both respondents and nonrespondents need to be available, which may come from a previous wave or an external data source (e.g. an administrative register or census data). In addition, survey agencies have also started to collect call record data for interviewer administered surveys comprising, for example, date, time and outcome of calls. Such data are a form of paradata (Couper, 1998, 2000; Couper and Lyberg, 2005) and have been identified as a key tool to investigate nonresponse bias by survey practitioners. This is usually undertaken by analysing fieldwork information that is related to both the survey outcome and the response mechanism (Groves, 2006; Groves and Heeringa, 2006).

A number of recent studies have started to monitor nonresponse bias during fieldwork. The German Labour Market and Social Security panel survey (PASS) (Trappmann *et al.*, 2013; Trappmann *et al.*, 2014) combines automatically generated graphical displays with knowledge of paradata to follow differences in contact, cooperation and response rates between waves during fieldwork. The R-indicator (Schouten *et al.*, 2009), a measure of representativeness of respondents compared to the whole sample, is also monitored. Although advocated by Schouten *et al*. (2009) as an appropriate measure to monitor the call process, the R-indicator does not seem to have appropriate properties for low response rates, which is the case at the beginning of the call process. Kreuter *et al.* (2010) assess nonresponse error and measurement error linking wave 1 of PASS to administrative data. The study population, however, is limited, as the authors

consider respondents living in one-person household who recently received unemployment benefits and are under 65 years of age. The errors are assessed in terms of root mean square error, bias and standard deviation across groups, with groups defined in terms of level of effort in the data collection. Benefit recipiency status, respondent's current employment status, age and citizenship are the survey items monitored. They conclude that additional fieldwork efforts reduce nonresponse bias in their particular analysis. Balance indicators have been developed by Lundquist and Särndal (2013) as a quadratic distance between respondents and full sample means using a suitable weighting matrix. These indicators, along with the R-indicator, are followed across call attempts in the Swedish Living Conditions Survey 2009. The authors also monitor the relative differences between calibrated and Horvitz-Thompson (unbiased) estimators across call attempts. The conclusion is that balance in the sample has not increased after 20 call attempts and the fieldwork strategy should be revised, for example, by stopping data collection after 10 contact attempts. Schouten *et al.* (2013) propose a theoretical framework for optimising quality of response on surveys based on quality measures used in recent studies such as response rate, R-indicator, coefficient of variation of response propensities and estimated nonresponse bias, along with paradata information. The authors advise on the choice of loss functions for scenarios defined in terms of survey variables, parameters of interest and estimators to be used. Wagner (2012) presents an extensive review on key survey quality indicators (e.g. response rate, coefficient of variation, R-indicator and fraction of missing information) according to their types (indicators based on response indicator only; based on response indicator and paradata/frame data; based on response indicator, paradata/frame data and survey data). The indicators are illustrated with figures and tables extracted from different survey reports and their weaknesses and strengths are discussed, but without focusing on practical implementation and the features of

each indicator based on one specific survey. The author suggests, however, that further research is required to identify efficient indicators to detect risk of nonresponse bias as each survey has its own features and it is very unlikely that one indicator alone will capture this issue.

This paper aims to monitor nonresponse bias during the field process using call record data to inform survey practice and is motivated by the question on how many calls may be needed to achieve a good level of representativity. The use of dissimilarity indices is proposed to assess nonresponse bias in cross-sectional or longitudinal surveys based on the knowledge of key variables (either from internal or external sources). The main rationale of the proposed method is to compare the true (fully observed, not subject to nonresponse) distribution with the observed (subject to nonresponse) distribution of key variables. Whilst dissimilarity indices have been employed for various purposes (e.g. Agresti, 2013), they have not been applied so far in the context of nonresponse bias monitoring.

The proposed methods are compared to other existing and commonly used methods to monitor nonresponse at the end of the fieldwork process. Both survey-specific indicators (e.g. response rate, R-indicator and coefficient of variation) as well as variable-specific indicators (e.g. nonresponse bias, partial R-indicator, partial coefficient of variation and dissimilarity indices) are considered.

An appealing approach is to monitor and assess nonresponse bias and associated indices through graphical displays and this is advocated here. Call record data play an important role in this type of analysis since they allow quality indicators to be followed across number of contact attempts, number of noncontacts, or day or week of the data collection period.

The study applies the proposed methods to data from waves 1 and 2 of Understanding Society, a large-scale longitudinal household study in the UK including rich paradata (McFall,

2013). The principles, however, can be applied to any survey with access to fully observed information on both respondents and nonrespondents, either from internal or external data sources, including both cross-sectional and longitudinal surveys.

The findings will guide survey practitioners in the development of adaptive and responsive designs (Groves and Heeringa, 2006; Lundquist and Särndal, 2013; Schouten *et al.*, 2013; Trappmann *et al.*, 2014; Laflamme, 2013), for example, to define when best to stop calling on a sample unit and to identify cases which should be followed up, if the aim is to minimise nonresponse bias (Heerwegh *et al.,* 2007; Lundquist and Särndal, 2013). The approaches proposed here can be used to identify areas for savings in time and staff resources that could be employed elsewhere to strengthen other data quality aspects of the survey.

2. Methods

A range of methodological approaches are proposed and reviewed to monitor nonresponse bias for categorical and continuous variables during fieldwork. For generality, consider the scenario where two data sources are available: data source 1, where the "true" values are known for all units in a specific survey (e.g. administrative data or data available from a previous wave in a longitudinal study) and data source 2, providing field process information (e.g. call status and interview outcome) for *all* (responding and nonresponding) units. For the particular application in this paper data source 1 is the previous wave and data source 2 is the current wave of a longitudinal study.

Data source 1 is regarded as the target population that we aim to estimate, implying that this is effectively the "true" distribution. The representativeness of respondents in data source 2 is assessed by comparing estimated distributions of categorical and continuous variables based

on information collected at data source 1 (but restricted to only those units identified as respondents in data source 2) with the corresponding "true" distributions obtained from the whole set of responding units in data source 1. Only the response indicators are from data source 2, which allows the survey "true" values to be available for responding and nonresponding units.

Response Rate and Bias

Response rates are commonly monitored during data collection as an initial step to assess survey nonresponse. Response rates below a specific target may indicate the need for initial nonresponse bias analysis (Groves and Peytcheva, 2008). This includes, for example, contrasting information on respondents and nonrespondents according to characteristics of the sampling frame, administrative registers or from previous waves (in a panel survey) and assessment of response rates by groups. Efforts to increase response rates can result in increased nonresponse bias (Groves, 2006; Merkle and Edelman, 2009), in particular if response rates are different for some subgroups of the population and a study variable is related to the variable defining the subgroups.

The use of paradata for such studies have increased over the past 10 years with common applications including monitoring of data collection efficiency, nonresponse bias diagnoses and prediction of response propensities. In these studies, the need for intervention is identified at the data collection stage and follow-up strategies can be implemented. This includes use of incentives to interviewer and potential respondents, prioritising hard to find cases, subsampling nonrespondents for further investigation aiming at characterising them and use such information at the estimation phase (weighting methods), comparison of reluctant and complete responders or

early and later responders in terms of demographic characteristics (Lynn *et al.*, 2002; Abraham *et al.*, 2006), among others.

The percent response rate (PRR) is given by

$$PRR = 100 \times \frac{m}{n}, \quad (1)$$

where m and n are the number of responding units and the number of eligible units in the survey, respectively. This can also be computed within subgroups of survey variables.

The estimated bias (B) and percent relative bias (PRB) of an estimator $\hat{\theta}_Z$ for a parameter of interest θ_Z for a variable Z is given by (Groves, 2006; Schouten *et al.*, 2013)

$$\mathrm{B}(\hat{\theta}_Z) = \hat{\theta}_Z - \theta_Z \quad (2)$$

and

$$\mathrm{PRB}(\hat{\theta}_Z) = 100 \times \frac{\hat{\theta}_Z - \theta_Z}{\theta_Z}. \quad (3)$$

Dissimilarity Measures

Whilst expressions (2) and (3) are attractive to survey practitioners because they are quick and simple to compute and straightforward to interpret, they are variable-specific indicators and, for categorical variables, these measures are only appropriate for a relatively small number of categories. For variables with a large number of categories, the two measures become difficult to monitor as they are computed for each category.

The main rationale for the dissimilarly methods advocated here to evaluate nonresponse bias is to compare the observed distribution subject to nonresponse with the true distribution across a range of survey variables, including both categorical and continuous variables. A

number of indices or statistics exist in the literature to compare two distributions but they have not been used in the context of nonresponse bias monitoring.

Key advantages of such measures are that they cope with categorical variables with large numbers of categories and a threshold is recommended in the literature to indicate non-negligible dissimilarity. The dissimilarity indices have the additional advantage of allowing comparability of several variables in the same graph rather than in individual plots, making nonresponse bias monitoring during fieldwork more efficient. They do not require additional computational effort to be computed in practice since they are either straightforward to implement or available in standard statistical software, e.g. R (R Core Team, 2014). In what follows, the use of dissimilarity measures is proposed for categorical and continuous variables.

The delta dissimilarity index Δ_Z (Agresti, 2013) for a variable Z with K categories is given by

$$\Delta_Z = \sum_{k=1}^{K} \left| \hat{\pi}_{Z,k} - \pi_{Z,k} \right| / 2, \qquad (4)$$

where $\hat{\pi}_{Z,k}$ is the observed (computed for survey variables in data source 1 based on respondents from data source 2) proportion in category k of survey variable Z and $\pi_{Z,k}$ is the corresponding expected (from data source 1) proportion. This index ranges from 0 to 1. The higher the delta index the more dissimilar is the estimated distribution to the true distribution. Agresti (2013) suggests that values between 0.02 and 0.03 indicate that the estimated distribution (e.g. observed proportions) follow the "true" distribution (e.g. expected proportions) quite closely. In survey practice very small delta indices would then indicate no or negligible nonresponse bias.

Other dissimilarity measures are also available in the literature following the same rationale as the delta index for the comparison of two distributions. The Kullback-Leibler index

L_Z (Kullback and Leibler, 1951; Kullback, 1987) is one example of such a measure and is given by

$$L_Z = \sum_{k=1}^{K} \pi_{Z,k} \log\left(\frac{\pi_{Z,k}}{\hat{\pi}_{Z,k}}\right), \quad (5)$$

where $\pi_{Z,k}$ and $\hat{\pi}_{Z,k}$ are as before. Another example is the chi-square statistic (Chernoff and Lehmann, 1954), a commonly applied measure to contrast the distributions of two categorical variables in terms of the differences between expected and observed values for each category. However, the delta index presents the advantage of having a threshold above which there is indication of dissimilarity between the distributions being compared (Agresti, 2013). Therefore, the delta index seems preferable to monitor categorical variables.

For continuous variables the Kolmogorov-Smirnov (K-S) statistic (Conover, 1971, pp. 295-314) may be used. It quantifies the maximum distance between the "true" (based on data source 1) and the estimated (computed for respondents in data source 2) distribution functions of a survey variable. In addition, the estimated functions can be graphically compared since the distribution estimation step is required to obtain the K-S statistic. More formally, the K-S statistic for two probability distributions is defined as

$$K_{n,n'} = \sup_Z \left|F_{1,n}(Z) - F_{2,n'}(Z)\right|, \quad (6)$$

where $F_{1,n}(Z)$ and $F_{2,n'}(Z)$ are the empirical (cumulative) distribution functions of variable Z based on data source 1 with n observations and data source 2 with n' observations, respectively, and *sup* is the *supremum* of the distances.

R-indicators and Partial R-indicators

An indicator that has been advocated in recent years as a nonresponse bias indicator alongside the traditional measurement of the nonresponse rate is the R-indicator (Schouten *et al.*, 2009), which is reviewed here for comparison to the dissimilarity measures above. The R-indicator is a representativeness indicator which measures the similarity between the respondents and the entire sample. For a population of size N, it is given by

$$R(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}\,, \qquad (7)$$

where ρ_i is the unknown response probability of unit i when it is sampled, $\bar{\rho}$ is the average population propensity and $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}$ is the standard deviation of the response probabilities.

The estimator $\hat{R}(\rho)$ of $R(\rho)$ for a sample of size n (Schouten *et al.*, 2009; 2011) is given by

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} I_i w_i (\hat{\rho}_i - \hat{\bar{\rho}})^2}, \qquad (8)$$

where I_i assumes value 1 when unit i is in the sample and 0, otherwise; w_i is the sampling weight (inverse of the sample inclusion probability) of unit i; $\hat{\rho}_i$ is the estimated response probability of sampled unit i and $\hat{\bar{\rho}} = (1/N)\sum_{i=1}^{N} I_i w_i \hat{\rho}_i$ is the weighted average of the estimated propensity probabilities.

The R-indicator varies from 0 to 1 with values close to 1 meaning the nonresponse mechanism is ignorable or, in other words, there is no systematic difference between response and the sample. The more the sample units differ with regard to the propensity to respond to the survey the smaller the R-indicator will be. The response propensities ρ_i in equation (7) are

unknown (even for population units) and, therefore, need to be estimated. In general, these estimated response propensities may be obtained by using the predictions of a response propensity model (e.g. a logistic regression model), which depends on a particular set of auxiliary variables. Because the R-indicator is not variable-specific and is simple to implement, it is an attractive indicator to compare representativeness across several cross-sectional surveys or several waves of a longitudinal study. For comparison of representativity across different surveys or across time points for the same survey, the same set of potential auxiliary variables should be available and used across all surveys or time points. It is recommended to include in models a limited number of explanatory variables, as the R-indicator showed to be sensitive to the number of variables included in the model (Schouten *et al*., 2009, p. 106).

Another point to consider is that small sample sizes tend to yield an unrealistic optimistic result for the R-indicator, since there will be no room for the response propensities to vary and the R-indicator will be large. For homogenous groups of respondents the R-indicator would be close to 1 potentially indicating a high representativity and hence high data quality. However, for the analysis of nonresponse bias across calls this may be a disadvantage since for the early stages (such as at the first or second call) the group responding may be quite homogenous leading to the R-indicator being close to one. This would then wrongly imply a high representativity, since no account has been taken of the sample size. Another potential complication of the R-indicator when applied during fieldwork is that a response propensity model would need to be fitted at every stage of data collection (see Schouten *et al.* (2009) for a detailed discussion on R-indicator features).

Unconditional and conditional partial R-indicators have also been developed (Schouten *et al.*, 2011; 2012) and measure the similarity between respondents and the entire sample for

subgroups of the population defined by known auxiliary variables. The unconditional partial R-indicator for variable Z with K categories is defined by

$$P_u(Z,\rho) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{K} N_k(\bar{\rho}_k - \bar{\rho})^2}\,, \qquad (9)$$

where $\bar{\rho}_k$ is the average response propensity for category k of variable Z and N_k is the population size of category k. It is assumed that Z is included in the vector X of auxiliary variables used in the response propensity model. The unconditional partial R-indicator ranges from 0 to $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}$, which has 0.5 as an upper bound. The larger $P_u(Z,\rho)$ the greater the response propensities variability between the subgroups defined by categories of Z and, therefore, the greater the lack of representativeness in the sample that is attributed to variable Z.

The estimator $\hat{P}_u(Z,\rho)$ of $P_u(Z,\rho)$ is

$$\hat{P}_u(Z,\rho) = \sqrt{\frac{1}{N}\sum_{k=1}^{K} \hat{N}_k(\hat{\bar{\rho}}_k - \hat{\bar{\rho}})^2}\,, \qquad (10)$$

where $\hat{N}_k = \sum_{i\in s_k} w_i$ is the estimated population size of category k of variable Z, $\hat{\bar{\rho}}_k = (1/\hat{N}_k)\sum_{i\in s_k} w_i\hat{\rho}_i$ and s_k is the set of sample units in category k of variable Z.

The partial R-indicator shares the same undesirable features described above for the R indicator, namely the requirement to fit successive models at each stage of data collection, the sensitivity to number of variables used in the model and sensitivity for small sample sizes. This last characteristic, in particular, is illustrated in Section 3 when the partial R-indicator shows a very good performance at the first one or two calls, then a decline in performance and then a slow improvement again.

Coefficient of Variation and Partial Coefficient of Variation

The (partial) R-indicators seem to be sensitive to the specification of the model, the choice of variables in the model and the sample size. Although they have shown to be informative for comparison of different runs of the same survey, they may be less valuable for monitoring nonresponse bias during fieldwork in a particular study, which is the aim here. To overcome some of these potential problems it seems advisable to account for the response rate in the estimator. The coefficient of variation (CV) of the response propensities accounts for the average response propensity and is an upper bound for the absolute nonresponse bias, as shown in Schouten *et al*. (2009). However, the CV has not been advocated or commonly used as a survey quality measure, apart from the example shown in Wagner (2012). In fact, the CV has only recently been suggested under a theoretical framework (Schouten *et al.*, 2013). Its expression is given by

$$CV(\rho) = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}}{\bar{\rho}}, \tag{11}$$

and its estimator is given by

$$\widehat{CV}(\rho) = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} I_i w_i \left(\hat{\rho}_i - \hat{\bar{\rho}}\right)^2}}{\hat{\bar{\rho}}}. \tag{12}$$

Similarly, the unconditional partial CV for variable Z is defined by

$$PCV_u(Z,\rho) = \frac{\sqrt{\frac{1}{N-1}\sum_{k=1}^{K} N_k(\bar{\rho}_k - \bar{\rho})^2}}{\bar{\rho}}, \tag{13}$$

and its estimator is given by

$$\widehat{PCV_u}(Z,\rho) = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{K} \hat{N}_k \left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}\right)^2}}{\hat{\bar{\rho}}}. \tag{14}$$

In the special case where only the variable Z is included in the response propensity model, expressions (11) and (13) are, in fact, the same.

Although, on first site, the partial CV and the delta index seem to be quite different measures, there is a close link between them. The CV is based on the Euclidean distance between the response probabilities and their average, whereas the delta index is based on the city-block distance. To illustrate this link, let us consider the case of one variable Z in the propensity model. Rewriting the delta index in expression (4), where m_k and n_k are the number of responding units and the sample size in category k of variable Z, respectively, m is the number of responding units for variable Z and n is the total sample size, we obtain

$$\Delta_Z = \sum_{k=1}^{K} \frac{|\hat{\pi}_{Z,k} - \pi_{Z,k}|}{2}$$

$$= \frac{1}{2} \sum_{k=1}^{K} \left| \frac{m_k}{m} - \frac{n_k}{n} \right|$$

$$= \frac{1}{2} \sum_{k=1}^{K} \frac{n_k}{n} \left| \frac{m_k/n_k}{m/n} - 1 \right|$$

$$= \frac{1}{2} \sum_{k=1}^{K} \frac{n_k}{n} \left| \frac{m_k/n_k - m/n}{m/n} \right|$$

$$= \frac{1}{2} \left(\frac{m}{n} \right)^{-1} \sum_{k=1}^{K} \frac{n_k}{n} \left| \frac{m_k}{n_k} - \frac{m}{n} \right|$$

$$= \frac{\frac{1}{n} \sum_{k=1}^{K} n_k |\bar{\rho}_k - \bar{\rho}|}{2\bar{\rho}}. \qquad (15)$$

As a result, one would expect both measures to perform similarly in the application and to lead to similar conclusions.

3. The Data

Understanding Society – the UK Household Longitudinal Study (UKHLS) (University of Essex, 2012; McFall, 2013) is a large scale multi-thematic survey that collects information on socioeconomic factors such as education, employment, health, income and behavioural and attitude indicators. The UKHLS sample size is approximately 40,000 responding households to wave 1. Its broad scope and its longitudinal aspect make the UKHLS an unusual and valuable source of information.

Besides being a rich source of socioeconomic information, the UKHLS also collects information on field process data (or paradata) for responding and nonresponding households at all survey waves allowing the monitoring of nonresponse across the fieldwork period. Paradata available in the UKHLS include interviewer observations and call record data such as date and time of individual and household interviews, outcome of the calls to a household or call status (no reply, contact made, appointment made, any interviewing done and any other status) and individual interview outcome. Data collection at a particular wave runs across 24 months and is mostly face-to-face via Computer Aided Personal Interview (CAPI).

3.1 Analysis Sample

In this work, survey data from wave 1, collected from January 2009 to March 2011, and call record data from wave 2, collected from January 2010 to March 2012, are considered. The analysis sample was obtained by linking responding individuals in wave 1 (50,994) to their corresponding call record data collected in wave 2. A total of 955 individuals who responded at wave 1 became ineligible at wave 2 and were excluded from the analysis. Ineligibility information was obtained from the individual interview outcome variable at wave 2 and includes

the following categories: lost CAPI interview, unknown eligibility, out of scope/non-interviewed household, temporary sample member – not original sample member/permanent sample member, withdrawn before field, other ineligible and dead.

Call record data in the UKHLS is strictly speaking available at household level only. However, information on date, time and outcome of each call is available at household level and date of interview is available at individual level, which allows recovery of the outcome on each call at the individual level. A total of 3,035 individuals presented insufficient information on call record data at wave 2 to allow individual date of interview to be linked to household date of call. Therefore, these individuals were excluded from the analysis.

The analysis sample, therefore, contains 47,004 individuals. It consists of all responding individuals in wave 1 who were eligible for the survey in wave 2 and who presented sufficient call record information collected during the fieldwork process in wave 2. Thus, the analysis is conditioned on individuals who responded at wave 1.

The maximum call sequence length (series of call statuses) is 30 call attempts, with the mean and median number of calls equal to 3 and 2, respectively. Only 4% of individuals had more than 10 contact attempts. The response rate for the analysis sample is 76.5%. Table 1 shows the distribution of individual interview outcome for the analysis sample.

[Table 1 about here]

The individual response indicator at each call equals to one if the call outcome is “any interviewing is done” and to zero, otherwise. As mentioned in Section 2, responding individuals at wave 1 are regarded as the target population. Also, all wave 1 responding individuals were attempted to be interviewed in wave 2, so the sampling weights are one when computing the

survey estimates at each call attempt. The original sampling weights would be required in the analysis if one wanted to make inference about characteristics of the UK population, which is the target population for the UKHLS. All analyses are conducted at the individual level.

3.2 Variable Selection

The variables chosen to be monitored in this work vary in terms of type (binary, categorical and continuous), stability over time (tendency to be stable or change across waves) and survey topic to cover some of the main themes investigated in the survey (demographic characteristics, employment, health and income). The variables monitored in this study are: existence of long-standing illness or impairment, sex and did paid work last week (all binary variables); general health (with five categories ranging from excellent to poor) and age group (with quintiles as cut-points); age continuous and gross pay per month in current job - last payment. Any item missing cases are excluded from the analysis for each variable separately.

In addition to the variables listed above, further variables are used to predict the response propensities for the R-indicator, the partial R-indicator, the CV and the partial CV: born in the UK (binary), highest educational qualification (with six categories) and the interviewer observation variable on presence of children in the address (with five categories). The sample used in the prediction of the response propensities consists of 46,476 individuals as only complete cases across all variables are considered. The criteria to choose those variables are small number of categories (for non-continuous variables) and likely to be available from external data sources (census and administrative registers).

4. Results

The methods presented in Section 2 are applied to the analysis sample in order to illustrate how they can be used to assess and monitor nonresponse bias via graphical tools. This will inform survey practitioners on fieldwork effort required to achieve representativeness in the sample and provide guidance for future implementation of the survey. The plots proposed monitor survey variables across number of call attempts, but they could also be implemented to monitor variables across number of noncontacts (or "no reply"), for example, since different types of nonresponse are recorded in the call outcome.

All results are displayed until call 10 since the pattern observed for the quantities being monitored remains the same for subsequent calls. In general, it is worth monitoring the results across all calls since it is perceivable that nonresponse bias could increase with further calls. The number of respondents is cumulative across call attempts, e.g. the quantities computed at call 2 include the respondents at call 1 and call 2. We start analysing the binary and categorical variables. First, the development of the response rate and estimated bias across calls are assessed.

Figures 1a and 1b show the response rate and relative bias per call number for age group and existence of long-standing illness or disability (binary), respectively. Firstly, it can be seen for both variables that, although the response rates increase with the number of contact attempts, the estimated proportions still differ from the corresponding true values even after 10 contact attempts, indicating persistent nonresponse bias. The percentage of people aged 63 years and over, for example, is overestimated by 10.8% after 10 call attempts, despite the fieldwork effort to obtain an overall response rate of 72% (Figure 1a). Similarly for the variable existence of long-standing illness or impairment, which shows relative bias of 5.1% after 10 contact attempts for those responding "yes" (Figure 1b). A nonresponse bias of 4.4% was also observed for

category “excellent” of variable general health, with a response rate of 69% (variable not shown). Nonresponse bias is not observed for call 10 or later for the other categorical variables monitored.

The second point to observe is that both graphs indicate that the relative biases stabilise after around 5 contact attempts for all categories in both variables. The same pattern is observed for the other categorical variables. This means that after 5 calls, although the response rate continuous to increase, almost no improvements are achieved in terms of nonresponse bias. These findings may signal potential for savings in survey costs if more effective fieldwork monitoring strategies are adopted, in contrast to the common practice of arbitrarily calling sample members based exclusively on response rate results.

[Figure 1 about here]

Dissimilarity measures, such as the Kullback-Leibler, delta and chi-square statistics, allow the display of all information contained in Figure 1, and even more for all binary and categorical variables, simultaneously in one graph. Since the three indices show the same pattern across call attempts (see Figure 2 for an example applied to the categorical variable general health), results are reported only for the delta index because of its straightforward interpretation.

[Figure2 about here]

Figure 3 shows that the delta index value is below 0.03 for all categorical variables (except age group) after 4 contact attempts, decreasing even further after that. The value 0.03 is the recommended threshold under which there is indication of no divergence between the true distribution and its corresponding estimate after a specific number of call attempts (Agresti,

2013). For age group, the index stabilises at around 0.04 after 10 contact attempts, indicating that some bias observed in Figure 1 is still present.

[Figure 3 about here]

To monitor estimates across call attempts for continuous variables, methods analogous to those described above are adopted. Figure 4 shows the Kolmogorov-Smirnov (K-S) statistic for variables age and last gross pay per month in the current job. Again, the K-S statistic stabilises after about 5 calls for both variables. Also, the density functions for variable age is estimated for each call number and compared to the corresponding true densities (Figure 5). It can be seen from Figure 5 that the age distribution after 10 contact attempts shows remarkable departure from the true age distribution, indicating underestimation of the percentages of young and overestimation of the percentages of old people after 10 contact attempts. This is in line with the findings from Figure 1a.

[Figures 4 and 5 about here]

Figure 6 shows the performance of the overall response rate, R-indicator (equation (8)) and CV (equation (12)) across call attempts for different choices of variables included in the prediction of response propensities. We start by analysing the results when including in the propensity models age group (the same pattern is observed when including age continuous or as a quadratic term for both the R-indicator and the CV). In call 1, the R-indicator achieves its highest value, 0.93, due to the overall response rate being below 10%, as anticipated in Sections 1 and 2. The R-indicator changes slightly from call 2 to 10, varying from 0.79 in call 2 to 0.81 in call 10, which indicates fairly good representativeness even in call 2 with overall response rate of

30%. The results based on the R indicator may therefore be misleading if not interpreted correctly. Using the R-indicator one may conclude that 1 call may be sufficient to reach a good level of representativeness. For further calls the R-indicator steadily increases with its maximum in the last call attempt, therefore, not indicating a natural level when further calls may not add much to representativeness. A relative measure accounting for the average response rate, such as the CV, is, therefore, more appropriate. The CV decreases across calls, ranging from 0.52 in call 1 to 0.11 in call 7 (remaining stable from call 7 to call 10), displaying a very similar pattern to the dissimilarity measures.

[Figure 6 about here]

In order to assess how sensitive the R-indicator and the CV are to different choices of explanatory variables, the variable age was excluded from the model (Figure 6). This yields an improvement in the R-indicator at a similar rate across calls, because in this example not conditioning on age implies less variability among the response probabilities. The same impact is observed on the CV.

Figures 7a and 7b show the partial R-indicator (equation (12)) with its upper bound and the partial CV (equation (14)), respectively, for each one of the variables considered for the delta index to allow direct comparisons: sex, age group, long-standing illness (binary), general health (5 categories) and did paid work last week (binary). Besides presenting the same limitations as the R-indicator, the partial R-indicator (Figure 7a) shows an undesirable feature by starting at a lower value at call 1 and increasing thereafter, as a consequence of not accounting for the response rate. Also, the partial R-indicators are small compared to the upper limit, indicating that the lack of representative attributed to each of the variables considered is small. This result is not

supported by Figure 1a. The partial CV (Figure 7b), on the other hand, shows patterns similar to that revealed by the delta index (Figure 3) for all variables, with age group standing out in terms of lack of representativeness. The delta index, however, is much easier to implement and does not require fitting a model at every call, which is necessary to obtain the CV and the partial CV.

[Figure 7 about here]

5. Concluding Remarks and Implications for Survey Practice

This paper presents and compares a range of approaches for monitoring and assessing nonresponse bias during the data collection process in a cross-sectional or longitudinal survey based on call record data and key variables from administrative data, register data or a previous wave. In particular, the paper proposes the use of dissimilarity indices which is a novel approach in the context of nonresponse bias analysis across calls. The main motivation of the methods is to compare the distributions of survey variables (subject to nonresponse) at each call attempt with the corresponding fully observed ("true") distributions. Specifically, the delta index, Kulback-Leibler, chi-square and K-S dissimilarity measures are investigated. For comparison, the paper reviews commonly used indicators including the response rate, relative bias, R-indicators and partial R-indicators, as well as a method more recently advocated, the coefficient of variation, and the partial coefficient of variation, which has not been used in this context before. Some advantages and disadvantages of the various methods are highlighted. Graphical representations to monitor new and existing indicators across call attempts are investigated to guide survey practitioners on when to apply interventions with the potential to adapt the survey design to save survey resources or to inform the reduction of nonresponse bias. A selection of survey variables from the UK Household Longitudinal Study, Understanding Society, is used to illustrate the

approach, which can be generalised to a larger set of survey variables. A key advantage is that the proposed methods can be quite easily implemented into existing routine monitoring, for example into those using dashboards (Laflamme, 2013).

The main findings are as follows:

1. The results show that survey estimates tend to stabilise after about 5 calls and, in some cases, at levels that depart significantly from the corresponding true values even after high response rates have been achieved.
2. Key advantages of the dissimilarity measures are that they are able to adequately capture nonresponse bias during the fieldwork process, they are simple to compute (without the need of fitting a model at every call and related modelling assumptions) and a threshold (e.g. delta index) can guide indications for the lack of representativeness. Representation of several variables in the same graph is possible, even for variables with large number of categories, which simplifies the assessment.
3. The coefficient of variation leads to very similar patterns and conclusions as the dissimilarity measures, supporting the findings. However, the CV relies on fitting a response propensity model at every call, requiring modelling assumptions and decisions on which variables to include in the nonresponse models.
4. The paper also highlights some potential limitations of extending the use of indicators commonly used in the literature for nonresponse bias monitoring to data collection monitoring. Although adequate for the use after data collection, R-indicators and partial R-indicators present undesirable features during fieldwork monitoring. Since the indicators do not account for the nonresponse rate, it makes them sensitive to low response rates and they

may even give misleading results. Hence, such methods are not advocated in the context of nonresponse bias monitoring during fieldwork.

5. Among the survey quality indicators discussed in this paper, it is shown that, the delta index, the relative bias and the partial CV are the only indicators able to quantify that the sample does not present appropriate level of representativeness (nonresponse bias) after a number of contact attempts.
6. One problem is that, generally, call record data are only available at the household level, whereas variables of most interest for nonresponse bias analysis are often at the individual level. This paper provides guidance to survey researchers and practitioners faced with similar issues. We overcome this problem by identifying individuals within households, their respective date of interview (if responded) and matching the date of interview to one of the call dates to obtain information on nonresponding individuals and call data, and then tracing the individuals back to their wave 1 data to obtain the fully observed information.

Although in this study only a number of variables are monitored, the principles of the methods can be easily extended to more general scenarios. Whilst this study assesses nonresponse bias with regard to the distribution of wave 1 variables, this work could be extended to assess nonresponse bias with regard to the current wave, if administrative or other external data sources obtained at the time of the survey are available. For an example of such an application, see Kreuter *et al.* (2010), who use wave 1 data of the PASS survey linked to administrative data.

This research has direct implications for survey practice. There is a growing body of research on responsive and adaptive survey designs targeting two crucial issues currently affecting surveys agencies: reduction of survey costs and improvement of data quality. By

providing guidance on which survey quality indicators are appropriate to monitor nonresponse bias during fieldwork, the methods proposed assist in the process of identifying the best time to implement interventions and to redefine the fieldwork strategy or to stop calling. For example, knowing that after 5 call attempts the distribution of key survey variables are already similar to the corresponding true distributions, resources could be redirected to enhance the representativeness of the survey, instead of continuing with the same strategy up to call 30. Although less attractive for a longitudinal study, where the aim often is to retain sample members for as long as possible in the survey, in a cross-sectional study, where longer-term attrition is not an issue, survey researchers could opt for stopping further calls and use the saved resources to enhance other data quality features of the survey. On the other hand, if it is known that the two distributions being compared are not similar enough after say 6 calls, subgroups underrepresented in the survey could be prioritised by changing the fieldwork strategy.

Funding

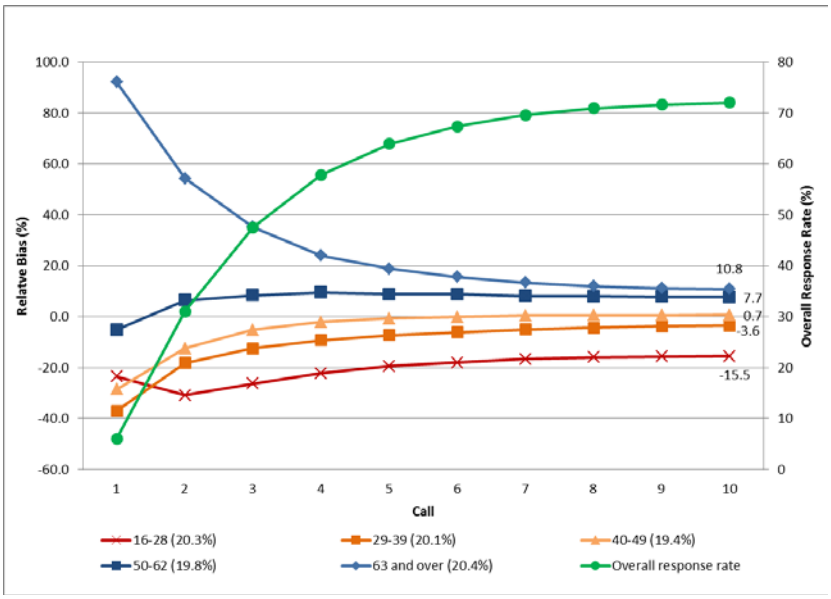
This work was supported by the research grant 'The use of paradata (field process data) in cross-sectional and longitudinal surveys' funded by the UK Economic and Social Research Council (ESRC) [ES/I018301/1] and the ESRC National Centre for Research Methods 2014-2019 (research workpackage 1)[ES/L008351/1].

Acknowledgements

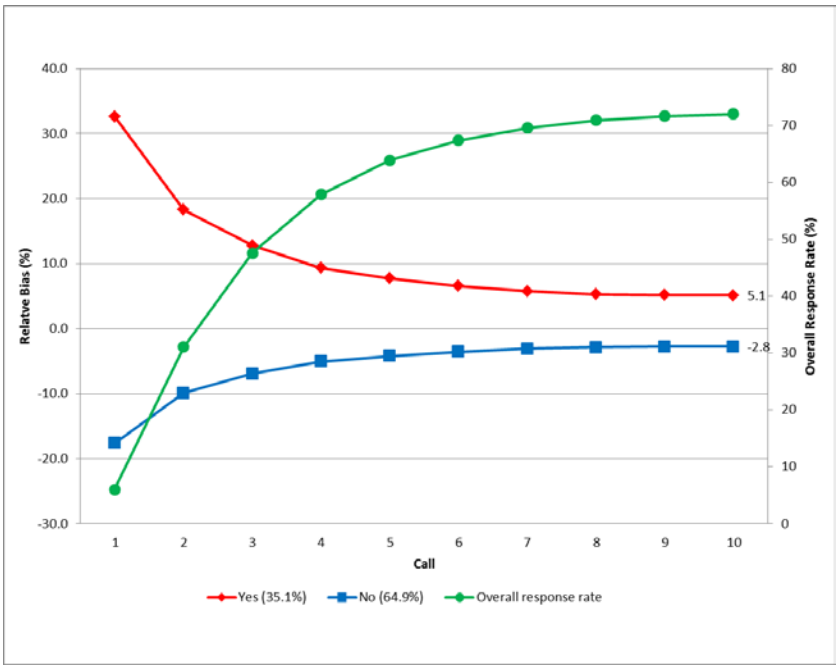
The authors thank Professor Nathalie Shlomo (University of Manchester) and Professor Li-Chun Zhang (University of Southampton and Statistics Norway) for valuable comments to earlier discussions of this work.

Table 1. Distribution of Individual Interview Outcome for the Analysis Sample

Individual Interview Outcome	Frequency	Percent
Full interview	34,387	73.2
Proxy interview	2,057	4.4
Refusal	671	1.4
Other non-interviewed	457	1.0
Moved	2	0.0
Ill/away during survey period	158	0.3
Too infirm/elderly	32	0.1
Language difficulties	11	0.0
Refusal/non-interviewed household	5,026	10.7
Language problems/non-interviewed household	35	0.1
Non-contact/non-interviewed household	4,163	8.9
Ill/away – non-interviewed household	5	0.0
Total	**47,004**	**100.0**



(a)



(b)

Figure 1. Relative Bias and Response Rate for Variables Age Group (n=46,004) and Long-standing Illness or Impairment (n=46,892) Across Call Number. (Figures in brackets are the "True" Values.)

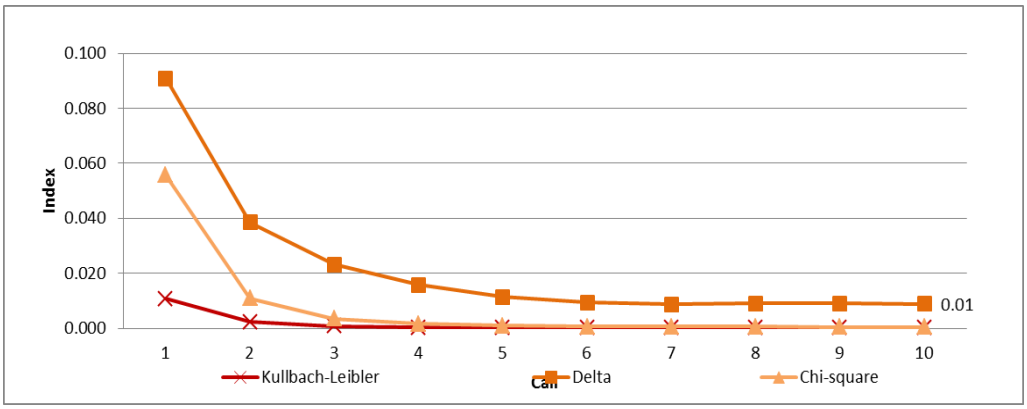


Figure 2. Dissimilarity Indices for Variable General Health with Five Categories.

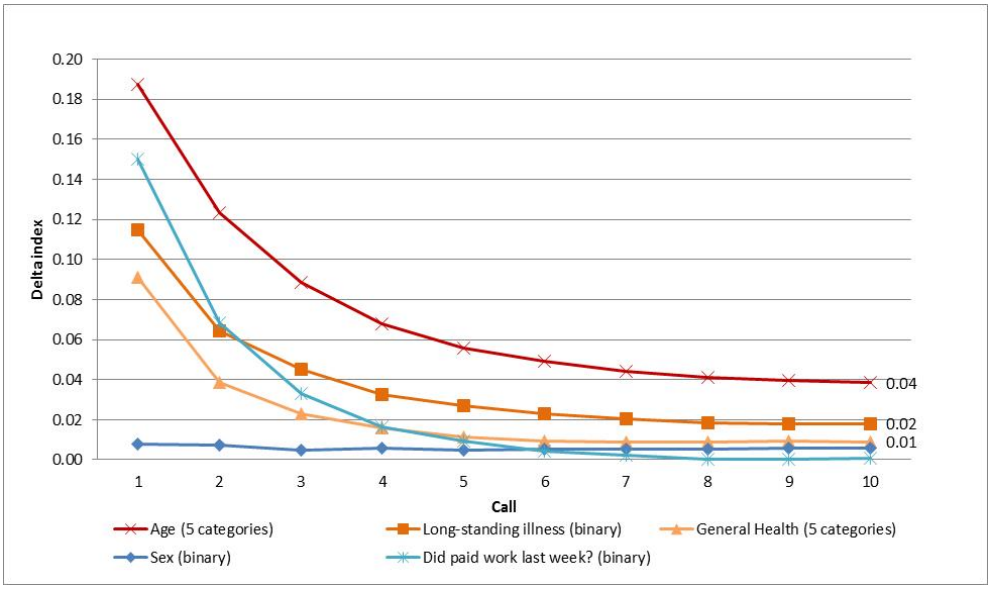


Figure 3. Delta Index for Binary and Categorical Variables.

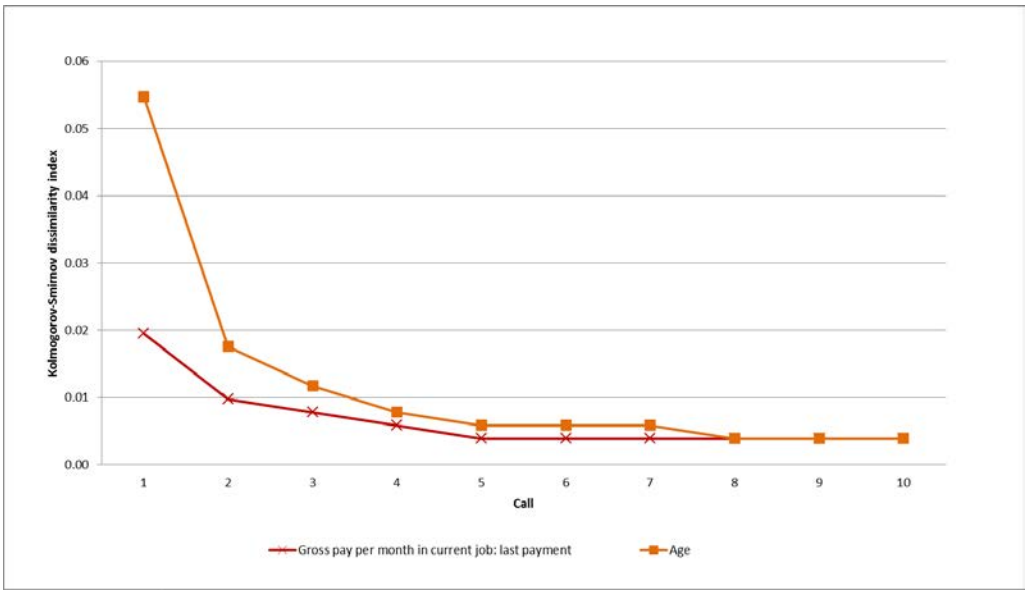


Figure 4. Kolmogorov-Smirnov measure for Continuous Variables.

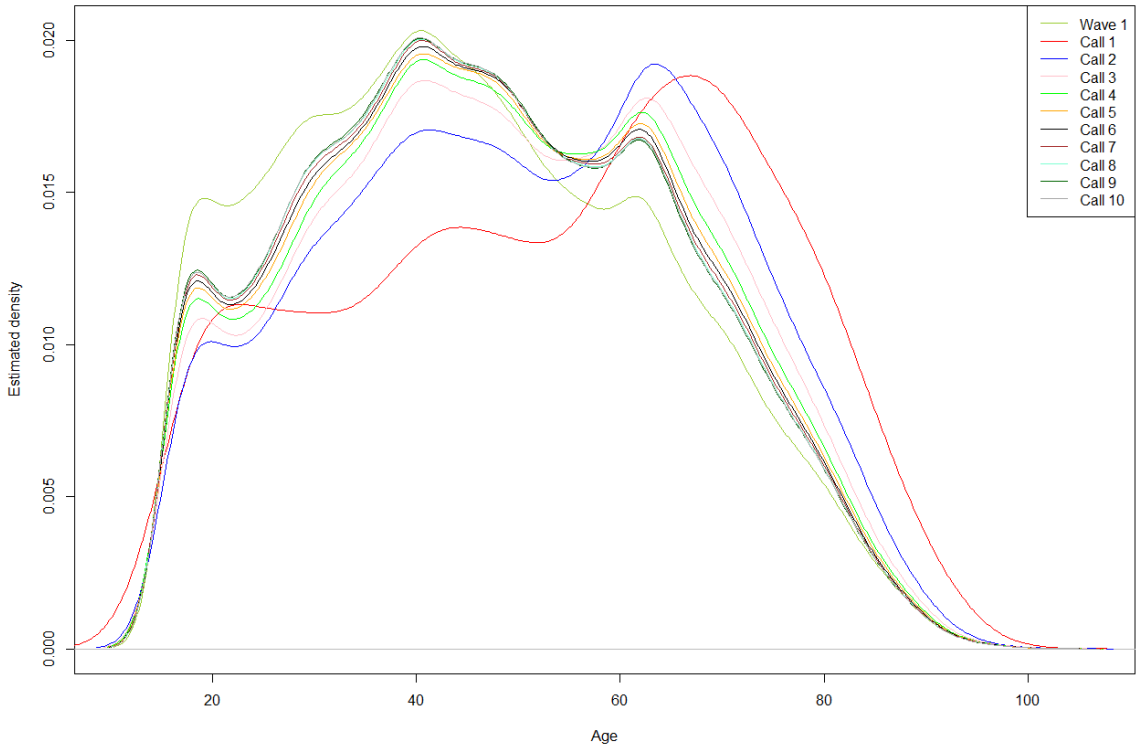


Figure 5. Estimated Density Functions for Variable Age (Continuous) Across Call Number.

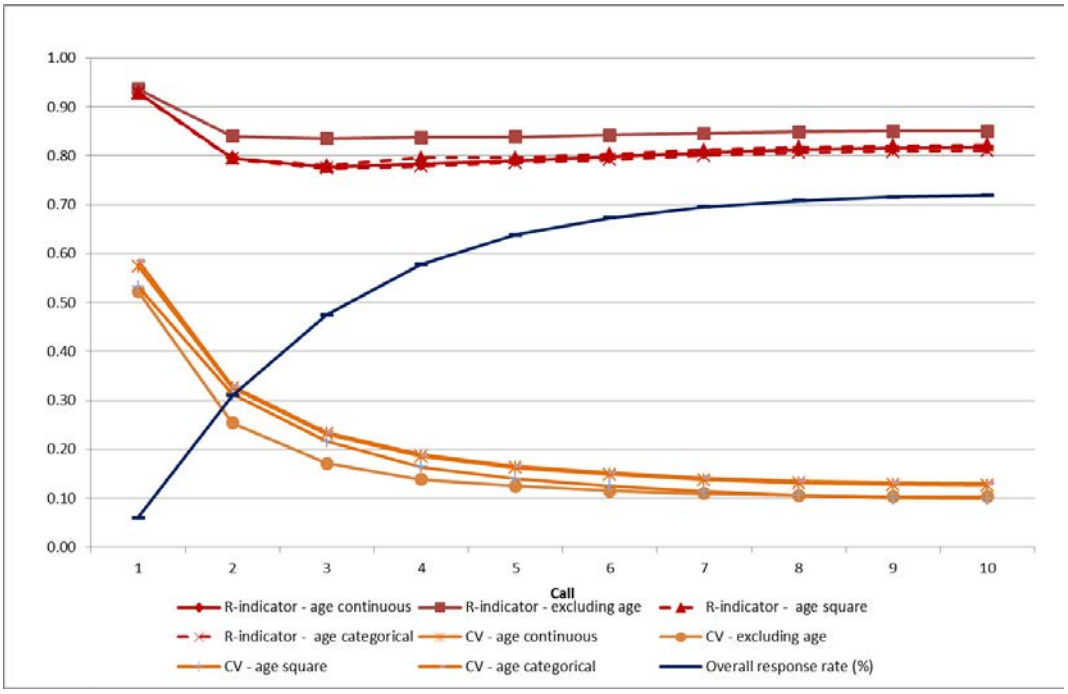
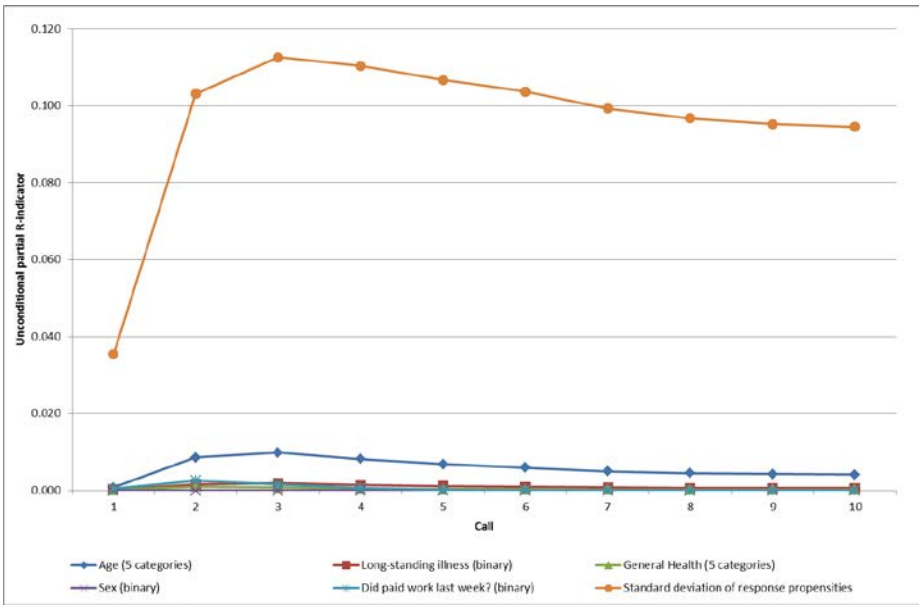
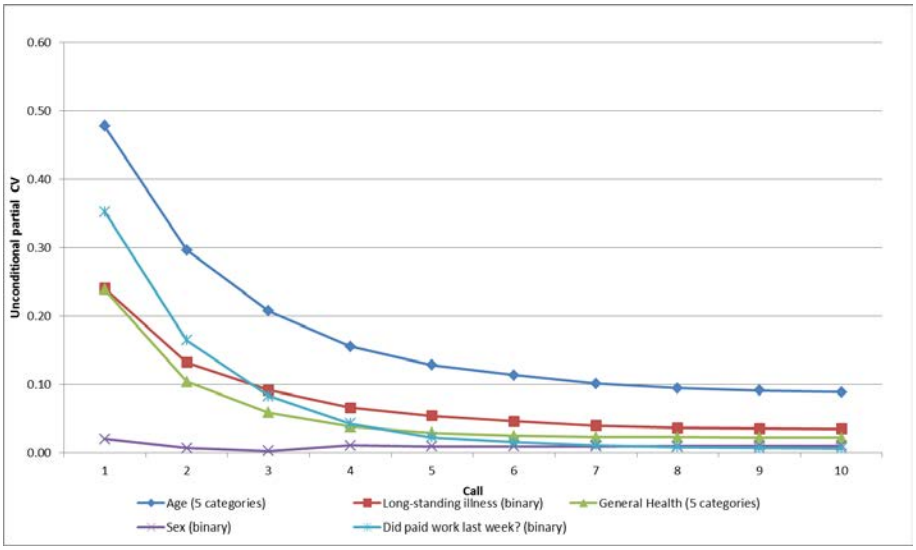


Figure 6. Overall response rate, R-indicator and Coefficient of Variation Across Call Number.



(a)



(b)

Figure 7. Unconditional Partial R-indicator and Unconditional Partial Coefficient of Variation for Binary and Categorical Variables.

6. References

Agresti, A. (2013) *Categorical Data Analysis*. Third Edition. New York: Wiley.

Abraham, K. G., Maitland, A., and Bianchi, S. M. (2006) Nonresponse in the American Time Use Survey: Who is missing from the data and how much does it matter? *Public Opinion Quarterly*, **70,** 676-703.

Chernoff, H. and Lehmann, E. L. (1954) The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. *The Annals of Mathematical Statistics*, **25**, 579–586.

Conover, W. J. (1971) *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Couper, M. P. (1998) Measuring survey quality in a CASIC environment. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 41-49.

Couper, M P. (2000) Usability Evaluation of Computer-Assisted Survey Instruments. *Social Science Computer Review*, **18**, 384–96.

Couper, M. P. and Lyberg, L. (2005) The Use of Paradata in Survey Research. *Proceedings of the International Statistical Institute Meetings*, 1–5.

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 (Statistics in Society)*, **176**, 251-269.

Groves, R. M. (2006) Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, **70,** 646-675.

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

Groves, R. M. and Heeringa, S. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society: Series A (Statistics in Society), **169**, 439-457.

Groves, R. M. and Peytcheva, E. (2008) The impact of non-response rates on non-response bias: a meta-analysis. *Public Opinion Quarterly*, **72**, 167–189.

Heerwegh, D., Abts, K. and Loosveldt, G. (2007) Minimizing Survey Refusal and Noncontact Rates: Do Our Efforts Pay Off? *Survey Research Methods*, **1**, 3–10.

Kreuter, F. (2013) *Improving Surveys with Paradata: Analytic Uses of Process Information*. Indianapolis: John Wiley & Sons.

Kreuter, F. Müller, G. and Trappmann, M. (2010) Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly*, **74**, 880 – 906.

Kullback, S. (1987) Letter to the Editor: The Kullback–Leibler distance. *The American Statistician*, **41**, 340–341.

Kullback, S. and Leibler, A. R. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22,** 79–86.

Laflamme, F. (2013) Using Paradata to Manage Responsive Collection Design. *International Workshop on Advances in Adaptive and Responsive Survey Design*, Heelen, Netherlands, 9-10 December.

Lundquist, P. and Särndal, C-E. (2013) Aspects of responsive designs with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, **29**, 557-582.

Lynn, P., Clarke, P., Martin, J. and Sturgis, P. (2002) The effects of extended interviewer efforts on non-response bias, in *Survey Nonresponse* (eds. R.M. Groves, D.A. Dillman, J. Eltinge and R.J.A. Little), pp. 135-147. New York: Wiley.

Merkle, D. M. and Edelman, M. (2009) An Experiment on Improving Response Rates and Its Unintended Impact on Survey Error. *Survey Practice* (March), 6–10.

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

R Core Team. (2014) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ .

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012) Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, **80**, 382-399.

Schouten, B., Calinescu, M. and Luiten, A. (2013) Optimizing quality of response through adaptive survey designs. *Survey Methodology*, **39**, 29-58.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the Representativeness of Survey Response. *Survey Methodology*, **35**, 101-113.

Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, **27**, 231-253.

Trappmann, M., Beste, J., Bethmann, A. and Müller, G. (2013) The PASS panel survey after six waves. *Journal for Labour Market Research*, **46**, 275-281.

Trappmann, M., Müller, G. and Kreuter, F. (2014) Introducing Adaptive Design Elements in the Panel Study "Labour Market and Social Security" (PASS). *4th Panel Survey Methods Workshop*, Ann Arbor, USA, 20-21 May.

University of Essex. (2012) *Institute for Social and Economic Research and National Centre for Social Research, Understanding Society: Wave 1-3, 2009-2012 [computer file]*. 4th Edition. Colchester, Essex: UK Data Archive [Distributor], December 2012. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-5.

Wagner, J.. (2012) A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, **76**, 555–75.

ONLINE APPENDIX

Table A1. Distribution of Study Variables in the Analysis Sample (sample size is 47,004).

Name	Description	Distribution	
a_sex	Sex	Male	21334
		Female	25670
a_health	Long-standing illness or impairment	Yes	16466
		No	30426
		Missing	112
a_ukborn	Were you born in the UK, that is in England, Scotland, Wales or Northern Ireland?	Yes	37978
		No	9012
		Missing	14
a_jbhas	Did paid work last week?	Yes	24732
		No	22186
		Missing	86
a_sf1	General health	Excellent	8723
		Very good	15073
		Good	13085
		Fair	6710
		Poor	3329
		Missing	84
age group (in years)	Derived from a_dvage using quintiles as cut-off points	16-28	9543
		29-39	9447
		40-49	9114
		50-62	9327
		63 and over	9573
b_child	Based on your observation, is it likely that this address contains one or more children aged under 10 including babies?	Definitely has a child/children aged under 10	4595
		Likely	3312
		Unlikely	7906
		Definitely does not have a child/children aged under 10	10676
		Cannot tell from observation	20196
		Missing	319

a_payg	Gross pay per month in current job – last payment (in Pounds)	Minimum	0.08
		First quartile	833.33
		Median	1500.00
		Mean	1829.00
		Third quartile	2400.00
		Maximum	15000.00
		Not applicable	24237
		Missing	4020
a_dvage	Age at the last birthday (in years)	Minimum	16
		First quartile	31
		Median	44
		Mean	45.71
		Third quartile	59
		Maximum	101