Do Coefficients of Variation of response propensities approximate non-response biases during survey data collection?

Jamie C. Moore
Institute for Social and Economic Research, University of Essex, Essex, UK.
Department of Social Statistics and Demography, University of Southampton, Southampton, UK. Email: moorej@essex.ac.uk

Gabriele B. Durrant
Department of Social Statistics and Demography, University of Southampton, Southampton, UK. Email: g.durrant@soton.ac.uk

Peter W.F. Smith
Department of Social Statistics and Demography, University of Southampton, Southampton, UK. Email: p.w.smith@soton.ac.uk

Abstract word count: 200

Main paper word count: 7114

Address for correspondence:
Institute for Social and Economic Research, University of Essex, Essex, UK. Email: moorej@essex.ac.uk

Summary

We evaluate the utility of Coefficients of Variation of response propensities (CVs) as measures of risks of survey variable non-response biases when monitoring survey data collection. CVs quantify variation in sample response propensities estimated given a set of auxiliary attribute covariates observed for all subjects. If auxiliary covariates and survey variables are correlated, low levels of propensity variation imply low bias risk. CVs can also be decomposed to measure associations between auxiliary covariates and propensity variation, informing collection method modifications and post-collection adjustments to improve dataset quality. Practitioners are interested in such approaches to managing bias risks, but risk indicator performance has received little attention. We describe relationships between CVs and expected biases and how they inform quality improvements during and post-data collection, expanding on previous work. Next, given auxiliary information from the concurrent 2011 UK census and details of interview attempts, we use CVs to quantify the representativeness of the UK Labour Force Survey dataset during data collection. Following this, we use survey data to evaluate inference based on CVs concerning survey variables with analogues measuring the same quantities among the auxiliary covariate set. Given our findings, we then offer advice on using CVs to monitor survey data collection.

Keywords: Non-response bias, representativeness indicators, adaptive survey designs, phase capacity, data collection efficiency savings.

1. Introduction

Methodologists no longer advocate only maximising response rates to minimise risks of survey variable non-response biases (Olson 2006; Kreuter 2013). Such biases are not easily estimated because non-respondents are not sampled, so response rates are appealing indirect measures of dataset quality. However, they have declined in the last 30 years (e.g. de Leeuw & de Heer 2002) and have been shown to be only weakly related to biases, due to differences between respondents and non-respondents even when response rates are high (Groves 2006; Groves & Peytcheva 2008). Instead, assessing variation in response across groups defined by subject attributes that are correlated with the survey variables is advised, including monitoring during data collection if interview attempt details exist. This can inform method modifications to target under-represented subgroups and reduce the risks of bias and / or minimise costs (adaptive strategies: Groves & Heeringa 2006; Wagner 2008; Peytchev et al. 2010). Practitioner interest in this more refined approach to managing survey dataset quality is increasing, but limited information on the performance of the proposed bias risk indicators restricts use. In this paper, we address this issue by using survey variable data to evaluate the performance of one set of indicators, Coefficients of Variation of response propensities (CVs), when monitoring data collection.

CVs and their counterparts, R indicators (together, representativeness indicators), are potentially valuable tools for assessing survey dataset quality (Schouten et al. 2012; see also section 2.1). Both indicators quantify variation in sample response propensities estimated given a set of auxiliary covariates observed for all subjects in the issued sample. If these covariates are correlated with the survey variables, low propensity variation (representativeness) implies low non-response bias risk. Overall indicators quantify dataset representativeness. Partial decompositions quantify propensity variation associated with the auxiliary covariates. Unconditional forms measure deviations from representativeness (a random sample), and conditional forms deviations from conditional representativeness (a random sample given the stratifying covariates). Approximate standard errors also exist, enabling statistical inference. When monitoring data collection, datasets with different response rates are often compared, for example to identify design phase capacity (PC) points after which further increases in quality are limited and methods should be modified or data collection ended (e.g. Groves & Heeringa 2006). In such scenarios, CVs have better properties than R indicators (Moore et al. 2018a). Another useful functionality is that CVs predict the maximal standardised non-response bias of survey variable means (Schouten et al. 2011), i.e. they measure dataset quality on a scale interpretable by practitioners.

The use of representativeness indicators in empirical scenarios is restricted by a lack of information on their performance, i.e. on how well they predict non-response biases. Schouten et al. (2016) report that high dataset representativeness reduces biases, but Nishimura et al. (2016; see also Beaumont et al. 2014) add a cautionary note, showing that excluding survey variable predictors from auxiliary covariate sets can cause biases to be under-estimated. This paucity of work reflects difficulties in estimating biases, although these may be reduced if a survey variable analogue measuring the same quantity exists among the auxiliary covariates (see below). A particular limitation is that CV performance when monitoring data collection is not known. For example, Moore et al. (2018a) study three UK social surveys, using linked census auxiliary covariates to compute CVs after each attempt to interview non-responding households (the call record). They identify PC points as when the CVs first fall within numeric thresholds either of previous call values, useful during data collection to inform current efforts, or of best values over the call record, useful after collection to inform future sampling. Given their findings, they argue that collection can be ended earlier than currently, substantially reducing the total number of calls made, with little effect on the risk of bias. However, such inferences are not evaluated: overall CVs and PC points are not compared to those computed given survey variables. Another issue is that PC points are not identified using inferential methods, an alternative to the numeric methods developed in other work on the topic (e.g. Rao et al. 2008; Wagner & Raghunathan, 2010; Lewis 2017).

Here, we address this knowledge gap by using survey data to evaluate CV based inference concerning the risk of non-response bias during data collection. We utilise a UK unique resource linking social survey responses and call records to census information (the 2011 Census Non-Response Link Study (CNRLS)). We study the Labour Force Survey (LFS) individual dataset component of this resource, extending work on the household version (Moore et al. 2018a) to the sample unit. We evaluate inferences about survey variables with an analogue measuring the same quantity in the fully observed auxiliary covariate set used to compute the CVs. As we show, partial CVs make predictions about the 'non-response biases' in these analogues that, if the same quantity is measured, also hold for the survey variables.

We begin by describing the derivation and interpretation of CVs, including for the first time their predictions about auxiliary covariate 'non-response biases'. Then, we outline how they can inform dataset improvements. We detail how partial CVs identify targets for collection method modifications, and also how for similar reasons they identify auxiliary covariate sets for use in post-collection bias adjustments, a previously unreported functionality. Moreover, we describe how CVs can be used to identify PC points, including introducing novel inferential methods.

Next, we monitor LFS data collection, by computing CVs given a census auxiliary covariate set at each call in the record. We also identify CV PC points, using both numeric thresholds (see earlier) and the fore-mentioned inferential methods. Then, we evaluate CV inferences about survey variables with auxiliary covariate analogues, making and testing the assumption that variable – covariate pairs measure the same quantities. First, we compute logistic regression based estimates of auxiliary covariate category standardised 'non-response biases' and identify PC points to compare to CV based inference. Second, to assess survey variable – auxiliary covariate analogue similarity, we compare category proportions for survey respondents given each data source. Utilising our findings, we then advise on how to use CVs to monitor data collection.

2. Methods

2.1. Coefficients of Variation of response propensities (CVs) and their use

2.1.1. Derivation

CVs measure sample-subset similarity in terms of variation in response propensities estimated given an auxiliary covariate set observed for all subjects (de Heij et al. 2015). The overall CV quantifies dataset representativeness, by dividing the propensity standard deviation by its mean: for sample size *n* and auxiliary covariate set ***x*** producing the propensity vector $p_{\boldsymbol{x}}$,

$$\widehat{CV}(p_{\boldsymbol{x}}) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{p}_i - \hat{\bar{p}}\right)^2}}{\hat{\bar{p}}}, \quad (1)$$

where $\hat{p}_i$ is the (estimated) response propensity of subject i, $\hat{\bar{p}}$ the average response propensity, and the numerator its standard deviation (*SD*). The less propensities differ the smaller the CV, and the greater dataset representativeness. Moore et al. (2018a) advise using CVs to monitor data collection instead of R-indicators (R = (1 – 2*SD*)) because dividing *SD* by $\hat{\bar{p}}$ means the resulting indictors are less likely to suggest high representativeness at early calls due to low propensity variation at low response rates (see also Schouten et al. 2009). Partial unconditional and conditional CVs (CV$_u$s and CV$_c$s) are derived from respectively the between and within variance decomposition components, and are bounded by the overall CV.

CV$_u$s quantify univariate associations between auxiliary covariates and propensity variation. The CV$_u$ for covariate Z with K categories is:

$$\widehat{CV}_u(Z, p_{\boldsymbol{x}}) = \frac{\sqrt{\frac{1}{n}\sum_{k=1}^{K} n_k(\hat{\bar{p}}_k - \hat{\bar{p}})^2}}{\hat{\bar{p}}}, \tag{2}$$

where n_k is the number of observations in category k and $\hat{\bar{p}}_k$ the mean response propensity in category k. Large values suggest substantial between category variability and non-representativeness associated with Z. Category CVs decompose and are bounded by covariate CVs. The CV$_u$ for category k of Z is:

$$\widehat{CV}_u(Z_k, p_{\boldsymbol{x}}) = \frac{\sqrt{\frac{n_k}{n}}(\hat{\bar{p}}_k - \hat{\bar{p}})}{\hat{\bar{p}}}. \tag{3}$$

Values can be positive or negative, implying respectively over- or under-representation.

CV$_c$s quantify associations between auxiliary covariates and propensity variation conditional on the other auxiliary covariates. The CV$_c$ for covariate Z is:

$$\widehat{CV}_c(Z, p_{\boldsymbol{x}}) = \frac{\sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i \in l}(p_i - \hat{\bar{p}}_l)^2}}{\hat{\bar{p}}}, \tag{4}$$

where $\hat{\bar{p}}_l$ is the mean propensity of the lth of L cells in a cross-classification of $\boldsymbol{x}$ excluding Z and $\boldsymbol{x}$ is the covariate subset for the propensity modelling. The CV$_c$ for category k of Z is:

$$\widehat{CV}_c(Z_k, p_{\boldsymbol{x}}) = \frac{\sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i \in l} h_i(p_i - \hat{\bar{p}}_l)^2}}{\hat{\bar{p}}}, \tag{5}$$

where h_i indicates whether subject i is in category k. Large CV$_c$s imply substantial solely attributable non-representativeness. In addition, adjustments to correct for biases caused by estimating propensities exist, as do approximate standard errors which when converted into 95% Confidence Intervals (CV ± 1.96 × standard error) enable inference regarding (comparative) representativeness (de Heij et al. 2015). Population level analysis is also possible by applying weights.

2.1.2. CV inferences about survey variable non-response biases

Overall CVs predict the maximum absolute standardised bias of survey variable means when non-response correlates maximally with the auxiliary covariates. Given an unknown auxiliary covariate set explaining response behaviour ($\aleph$), the Horvitz-Thompson estimate of the bias of a survey variable is approximated by the covariance between sample response propensities and the survey variable divided by mean response propensity (Bethlehem 1988). This value is standardised by dividing by the survey variable sample standard deviation ($S(y)$, for variable y with response mean $\hat{\bar{y}}_r$). By replacing the numerator covariance with its absolute maxima, which by the Cauchy Schwartz inequality is the product of the two variables' standard deviations, the maximum absolute standardised bias is estimated. The overall CV approximates this if the auxiliary covariates $\aleph$ can be replaced by the utilised set $\boldsymbol{x}$ (de Heij et al. 2015), e.g.

$$\frac{Bias(\hat{\bar{y}}_r)}{S(y)} = \frac{Cov(y, p_{\aleph})}{\widehat{\overline{p_{\aleph}}}\, S(y)} = \frac{Cov(y, p_x)}{\hat{\bar{p}}\, S(y)} \leq \frac{SD\ S(y)}{\hat{\bar{p}}\, S(y)} = \frac{SD}{\hat{\bar{p}}} = \widehat{CV}(p_{\boldsymbol{x}}). \tag{6}$$

Whether auxiliary covariate set $\aleph$ can be replaced by set $\boldsymbol{x}$ is untestable. In practice, including correlates of both response propensities and survey variables is essential, or biases may be under-estimated (see 'Introduction'). We note that another indicator studied by Nishimura et al. (2016), the survey variable absolute maximum bias (= $SD\ S(y)/\hat{\bar{p}}$), is derived similarly given $S(y)$ (Schouten et al. 2009).

In contrast, partial CV predictions about auxiliary covariate (analogue) 'non-response biases' are not described in the literature. In terms of equation (6), response propensity – auxiliary covariate covariance is maximal. Hence, as they are derived from the between component of the variance decomposition (see section 2.1.1), CV$_u$s provide inferences about covariate category (focal vs. others combined) standardised mean biases.

For two-category covariates, the covariate CV$_u$ (equation (2)) should approximate the absolute value of this bias (which in this case is independent of the focal category). For multi-category covariates, *K* combinations of the focal category vs. the others exist. For these, the covariate CV$_u$ should approximate the maximum absolute value of the different biases that can be computed, a value that would be obtained if the observed degree of propensity variation were due to all category deviations from expected being identical except for the focal category.

Category CV$_u$s (equation (4)) concern only the deviation of the focal category from expected. Hence, they should approximate the minimum category bias, with under-estimation less when, due to category size or the deviation, its contribution to the covariate inequality is large.

To summarise,

$$\widehat{CV}_u(Z_k, p_{\boldsymbol{x}}) \leq \frac{|Bias(\hat{\bar{Z}}_{k\,r})|}{S(Z_k)} \leq \widehat{CV}_u(Z, p_{\boldsymbol{x}}), \tag{7}$$

where $S(Z_k)$ is the standard deviation of the dummy variable indicating membership of category k, and the upper bound is attained if $K = 2$.

CV$_c$s are derived from the within component of the variance decomposition (see section 2.1.1). Hence, they make similar predictions about auxiliary covariate category (focal vs. others) absolute standardised conditional mean biases (i.e. those remaining after conditioning given the other covariates). Therefore,

$$\widehat{CV}_c(Z_k, p_{\boldsymbol{x}}) \leq \frac{\left|Bias(\hat{\bar{Z}}_{k\,r})_c\right|}{S(Z_k)} \leq \widehat{CV}_c(Z, p_{\boldsymbol{x}}), \tag{8}$$

where $Bias\left(\hat{\bar{Z}}_{k\,r}\right)_c$ is the focal category conditional bias, and again the upper bound is attained if $K = 2$.

Inequalities (7) and (8) represent a further functionality of CVs of use when assessing dataset quality: if auxiliary covariate analogues measure the same quantities, partial CVs provide inferences about survey variable non-response biases (see also sections 2.2.3. and 2.2.4).

2.1.3. Using CVs to inform dataset quality improvements

Regarding modifications of collection method, CVs can be couched in terms of missing data mechanisms (Schouten et al. 2012). Overall CVs quantify deviations in response from missing completely at random (MCAR) given the auxiliary covariate set. CV_us quantify deviations from MCAR with respect to a given auxiliary covariate (category), and CV_cs similar deviations from missing at random (MAR), i.e. the extent to which response is not missing at random (NMAR) given the other auxiliary covariates. Hence, CV_us identify under-represented groups to target. CV_cs ensure efficient targeting: non-significance implies an impact also associated with other auxiliary covariates. The suggested strategy is to target categories with significant CV_cs and some with significant CV_us only if (non-significant) CV_cs indicate correlations exist with other categories (Schouten & Shlomo 2017).

Similar arguments underlie why CVs can also inform post-collection non-response bias adjustments. Non-survey variable specific methods, including inverse response propensity weighting (e.g. Roberts et al. 1987), often assume responses are MAR given an auxiliary covariate set explaining response behaviour. To identify such sets, Särndal & Lundström (2010; see also Särndal 2011) use the Coefficient of Variation of the weights as a quality measure (Lundquist & Särndal 2013 and Särndal & Lundquist 2014 also similarly derive 'balance' indicators for assessing dataset quality). If the weights or propensities are similarly estimated (weighting often uses an identity link, in contrast to the logistic link generally used for propensities), or the sample size is large, this measure is equivalent to the overall CV: dividing the standard deviation of a set of inverse values by their mean is equal to the same calculation using the raw values (see also Schouten et al. 2016). Given this, partial CVs can be used to identify auxiliary covariates to include in the covariate sets used in weighting adjustments. CV_cs quantify inequalities after adjustment assuming MAR given the other auxiliary covariates, so if the covariates with large values are excluded from such sets their impacts will not be addressed (we note here that the included covariates should also be correlated with the survey variables, or adjusted variable variances will be inflated: Little & Vartivarian 2005). In fact, CV_cs can be used when identifying modification targets to statistically select covariate set members. In contrast, Särndal & Lundström's methods, comparing all possible sets or covariate selection, use arbitrary thresholds for accepting more complex sets. This is the first time this functionality of CVs has been described.

2.1.4. Using CVs to identify phase capacity (PC) points

Design phase capacity (PC) points are points in the data collection process after which further quality increases are limited and methods should be modified or collection ended (Groves & Heeringa 2006). Moore et al. (2018a) use CVs to identify PC points in household call records in three UK social surveys (including the LFS, whose individual level dataset is studied in this paper). They identify overall CV points and CV_u points for auxiliary covariates and under-represented categories: the former can be used to identify when to end collection, while impacts measured by CV_us are modification targets, potentially separately. They use numeric methods, specifically two rules that reflect whether identification is during collection (informing current efforts) or after (informing future sampling): i) if the CVs imply quality decreases or are within threshold *a* of the previous call CV ('during'); and ii) if CVs imply quality decreases or are within *a* of the best call record CV ('after'). No information existed on call costs or other methods, precluding optimising data collection given such alternatives using, for instance, the methods of Schouten et al. (2013); this is also the case for the dataset in this paper. Different thresholds *a* give comparable results. It should be noted that category (covariate) CV_us are decompositions, so PC points should be earlier than or similar to those given covariate (overall) CVs, although this may not always hold as the latter combine (different) multiple inequalities.

An alternative to using numeric methods to identify PC points are inferential methods. Most research seeks to identify points given changes in non-response adjusted survey variables over calls (Rao et al. 2008; Wagner & Raghunathan 2010; Lewis 2017). Tests assess whether variable differences differ from zero, accounting in adjustment method specific ways for dataset dependencies caused by early call responses also being in later call datasets, so are not usable with CVs. CVs are the focus of Schouten et al. (2016), who to assess the representativeness – survey variable bias relationships develop a rank test that uses partial CVs given different auxiliary covariate sets and auxiliary covariate biases. This can be used to identify PC points, but only from multiple covariate CVs. Concerning identifying single CV points, ignoring dataset dependencies we suggest that one approach is to use CV 95% CIs. As with numeric methods, different rules can be constructed to reflect whether points are identified during or after data collection. A PC point is identified during collection if the CVs are non-significant (i.e. the 95% CIs include zero), imply quality decreases or the 95% CIs overlap the previous call CIs. An PC point is identified after collection if the CVs are non-significant, imply quality decreases or the 95% CIs overlap that for the call with the best CV. We use these rules for the first time below.

We note that when using inferential methods in empirical scenarios, significance levels need consideration. The CV CI widths decline as response rates increase (Moore et al. 2018a), so unless such levels are adjusted, the statistical power to identify CV differences will vary over calls (see Lewis 2017 for discussion of similar with non CV based tests). This is perhaps a reason to use numeric methods: another is when decisions are optimal before CV parity, due to, for example, the costs of the alternative data collection methods. In the work in this paper though, a single significance level is not an issue: we evaluate CV performance by comparing CV PC points with those based on estimated bias (whose CIs similarly change: see 'Results').

2.2. Evaluating CV based inference for survey variables with auxiliary covariate analogues

2.2.1. The Labour Force Survey (LFS) dataset

The Office for National Statistics 2011 CNRLS links January to July 2011 UK social survey samples and their survey responses to their 27th March 2011 census records, providing attribute information whether they are interviewed or not (Parry-Langdon 2011). Linkage is via subject address and personal detail (name, gender, date of birth) matching. Survey call records are also appended. Our focus here, the LFS, samples English and Welsh individuals aged over 15 on labour market topics (see ONS 2011). Simple random sampling of households (HHs) is used. ONS operatives seek to interview all HH occupants. Most interviews are face to face, but a telephone interview can be chosen (see also below). The LFS is longitudinal, but we consider wave one subjects only to avoid sample attrition effects. For this wave, 96.9% of HHs and 93.3% of subjects are linked to census records (Table 1). Hence, we can study the majority of the sample using (self-reported) census responses (see ONS 2014) which reflect their attributes at the time (though we cannot rule out biases without non-linked subject data: Moore et al. 2018a). The call record data detail outcomes of calls to HHs (up to 20), and do not exist for telephone contacted HHs and some others (29.8% of the sample; see also below). Most HH members are interviewed at the same call. However, in around 1% of HHs, two members are interviewed at different calls. For these, we use the interview order to assign members to calls.

In our analyses, we consider eight survey variable – census auxiliary covariate analogue pairs (Table 2). All impact on LFS response propensities (Durrant & Steele 2009; Steele & Durrant 2011; Durrant et al. 2010, 2011, 2013) and are likely to be associated with other survey variables. ‘Tenure’ is a HH response, ‘HH structure’ a derived response, ‘Located in London/SE’ a geographic identifier, and the others individual responses. For a number of subjects, some responses are missing. This can reflect item non-response, but often is due to statistical disclosure control or, as with LFS subjects aged over 64, not being asked some items. Often, two or more responses are missing, so to minimise correlations we exclude these subjects from the analysed dataset. However, 'Age', 'Gender', 'Tenure' and 'HH structure' remaining item non-responders are so few that disclosure issues arise. Hence, we also exclude them, so that variables / covariates lack No response (NR) categories. We exclude 24% of the sample due to missing responses. Use of the methods below shows that excluding these subjects and those without call records (see previously) from the dataset causes under-representation of those in owned HHs or Aged '16 to 27' compared to the sample (results not shown): we outline likely impacts on findings in section 3.1.1. Given these exclusions, 'Gender', 'Tenure' and 'Located in London/SE' have two categories. The other variables / covariates are multi-category. The analysed dataset contains 21150 subjects in 11491 HHs. The final survey response rate is 58.6%; 13.9% of subjects refuse interviews and 27.5% are not contacted.

2.2.2. Quantifying LFS dataset representativeness and identifying PC points

We quantify LFS dataset representativeness by computing CVs from response propensities estimated by using a logistic regression model with as main effects the census auxiliary covariates listed in Table 2. At each call in the record, we compute bias adjusted overall CVs, auxiliary covariate partial CVs, and CV 95% CIs. We do not conduct population level inference as some survey subjects are not studied, so the supplied weights are not useful. We compute the CVs using the R code of De Heij et al. (2015; see www.risq-project.eu). We then identify CV PC points. We identify overall CV points, and CV_u points for auxiliary covariates and selected under-represented categories, using the numeric and inferential 'during' and 'after' collection identification rules described in section 2.1.4. For the numeric method points, we use a threshold *a* of ±0.02: others give comparable results (not shown).

2.2.3. Comparisons with census auxiliary covariate category 'non-response biases'

We evaluate CV based inference about survey variables with auxiliary covariate analogues by first computing logistic regression based estimates of census auxiliary covariate standardised 'non-response biases' for comparison. We describe CV predictions about category 'non-response biases' in section 2.1.2. To evaluate them, for the three two-category covariates and the selected under-represented multi-category covariate categories, we code new binary covariates such that

$$y_i = \begin{cases} 1, & \text{if subject } i \text{ is in the category of interest} \\ 0, & \text{if subject } i \text{ is not in the category} \end{cases}$$

where $i = 1,...,n$. We let r_i be the response indicator for subject i at a given call, with $r_i = 0$ indicating that they have not responded to the survey and $r_i = 1$ that they have. Next, at each call in the record, we estimate non-respondent – respondent differences in the log-odds of category membership. We fit two statistical models. Model A estimates overall differences:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 r_i, \tag{13}$$

where $\pi_i = \Pr(y_i = 1|\, r_i)$ is the category membership probability, β_0 the non-respondent log-odds of membership and β_1 the β_0 – respondent log-odds difference. Model B estimates differences conditional on the auxiliary covariate set $\boldsymbol{d}_i$ (set $\boldsymbol{x}$ minus the covariate underlying y_i):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 r_i + \boldsymbol{\beta}^T \boldsymbol{d}_i, \tag{14}$$

where $\boldsymbol{\beta}$ is a vector of coefficients. In model B, as y_i and r_i are binary, a β_1 of zero implies response with regard to a category is MAR given the auxiliary covariates. Non-zero values quantify the extent to which it is NMAR (Barbosa 2014). This provides similar information to a $\mathrm{CV_c}$ (see section 2.1.3). In model A, β_1 quantifies the deviation from MCAR, providing similar information to a $\mathrm{CV_u}$. Then, from parameter estimates, we compute standardised 'non-response biases' for the categories of y_i as:

$$\mathrm{Bias}(\hat{\bar{y}}_r) = \frac{\frac{m}{n}(\bar{\pi}_r - \bar{\pi}_{nr})}{S_{\bar{\pi}_s}}, \tag{15}$$

where m is the number of non-respondents, $\bar{\pi}_r$ the respondent category membership probability, $\bar{\pi}_{nr}$ the non-respondent probability, and $S_{\bar{\pi}_s}$ the sample probability standard deviation (see Groves & Couper 1998). With model A, we back-transform parameter estimates to obtain $\bar{\pi}_r$ and $\bar{\pi}_{nr}$. Model B estimates are conditional, so we compute marginal category membership probabilities to obtain $\bar{\pi}_r$ and $\bar{\pi}_{nr}$, by using Hastie's (1992) 'safe prediction' method in the R package 'effects 3.1.2' (Fox 2003; Fox & Hong 2009). To obtain $S_{\bar{\pi}_s}$ we fit a null model and use the delta method (Oehlert 1992) in the R package 'msm 1.6.4' (Jackson 2011).

We also identify overall (model A) bias PC points to compare to CV points, by using the same methods (see section 2.1.4). We again utilise the delta method to estimate standard errors and 95% Cis for the bias. Concerning predictions, covariate level $\mathrm{CV_u}$ points for two-category CVs and bias points should correspond (see section 2.1.2). Given contributions to covariate inequalities, similarities between multi-category covariate category $\mathrm{CV_u}$ and bias points should also exist.

2.2.4. Survey variable – census auxiliary covariate analogue similarity

To assess survey variable – census auxiliary covariate analogue similarity, for studied categories we compute survey respondent proportions given each data source at each call. We compare values graphically and by using Z tests for independent sample proportions.

3. Results

3.1. LFS dataset representativeness and PC points

3.1.1. Response rate development and CVs

LFS responses accumulate at a decreasing rate over the call record, with minimal increases after call 9 and none after call 17 (Fig. 1). The overall CVs decrease, suggesting increased representativeness, at a declining rate. The corresponding 95% CIs (see Table 1 in the on-line appendix), which decrease in width over the call record (as other CV intervals also tend to do), all exclude zero, implying respondents are always significantly non-representative of the sample.

We report the partial CVs in Figures 2 & 3 and the corresponding 95% CIs in Tables 1 to 4 in the on-line appendix. The 95% CIs of most auxiliary covariate unconditional CV (CV_u) exclude zero, implying significant associated inequalities. The 'Located in London/SE' CV_us begin as the largest, decrease at a declining rate to call six, then increase slightly. The 'HH structure' CV_us increase to call four, then decrease slightly, and are largest in the final dataset. Five covariates exhibit smaller, similar final inequalities. The 'Age' and 'Activity last week' CV_us decrease at declining rates. The 'Ethnicity' CV_us decrease slightly. The 'Qualifications' CV_us decrease to call two, then increase slightly. The 'Tenure' CV_us begin non-significant, increase to call two, then decrease slightly. The 'Gender' CV_us are smallest of all, increase slightly, and are non-significant to call five.

The category CV_us suggest 'Located in London/SE' and 'Age' inequalities are due to under-representation of London/SE and subjects aged under 40, although many are eventually interviewed. The 'Activity last week' inequality reflects similar Employed under-representation and increasing Student under-representation. The 'HH structure' inequality is due to initial under-representation of Single adult and Single adult with children HHs, but the latter impact declines and Other HH becomes under-represented. The 'Ethnicity' inequality reflects under-representation of Asian, Other and NR, and increasing under-representation of Mixed and Chinese. The ‘Qualifications’ inequality is due to initial under-representation of NVQ4+, NVQ3 and NR, but the first two impacts decline and None becomes under-represented. The ‘Tenure’ and ‘Gender’ inequalities reflect under-representation of Not owned HHs and Males.

The conditional CVs (CV_cs) suggest some of these impacts are independent. Only the ‘Gender’ and ‘Tenure’ CV_cs are non-significant. Some ‘Qualifications’ and ‘HH structure’ CV_cs are larger than the CV_us, implying greater inequalities. The ‘Located in London/SE’, ‘Age’ and ‘Activity last week’ CV_cs are smaller, suggesting inequalities partly correlated with the other auxiliary covariates. Of the under-represented categories, Student, NVQ3, Not owned HH, Male and most ‘Ethnicity’ impacts disappear: the category CV_cs are non-significant. The London/SE, Employed, Mixed ethnicity, ‘Qualifications’ None and NR, ‘Age’ and ‘HH structure’ impacts do not. Many such impacts exist in the HH dataset, putatively due to groups being less contactable (Moore et al. 2018a). This likely also holds for (some of) those newly identified here. The Employed and not owned HH impacts are respectively increased and reduced by including excluded subjects (those missing multiple responses etc.) in the dataset (see also section 2.2.1). Regarding improving datasets, categories with significant CV_cs are method modification targets (see also section 2.1.3.). Some with significant CV_us only should also be included if their impacts may be correlated with those of other categories: for instance, Students and Not owned HHs. Similarly, all covariates except ‘Gender’ and ‘Tenure’ should be included in auxiliary covariate sets used in post-collection bias adjustments.

3.1.2. CV PC points

The numeric method overall CV PC points using the 'during' and 'after' rules are at calls four and five respectively (Table 3). As expected, since CV_us are decompositions, most auxiliary covariate CV_u points are at the same calls or earlier. The 'Gender', 'Tenure' and 'HH structure' 'during' and 'after' points are at calls two and one respectively, and similar the 'Qualifications' points at calls three and two, because, although the CV_u minima are at the earlier calls, the 'during' rule only detects later increases. The 'Ethnicity' points are at call two, the 'Located in London/SE' and 'Age' points at call four, and the 'Activity last week' points at call five. We identify (multi-category) auxiliary covariate category CV_u points for the under-represented categories 'Age' 28 to 39, 'Activity last week' Employed, 'Ethnicity' Asian, and 'HH structure' Single adult and No qualifications. These points are earlier than the CV_u and overall CV points, again as expected. The 28 to 39 and Employed 'during' and 'after' points are at calls three and four respectively, due to later CV decreases detected by the 'after' rule. For the others, the 'during' points are one call later than the 'after' points (calls one and two), again due to the former rule not detecting CV minima.

The inferential method PC points follow similar patterns. The overall CV points are at call five. The 'Gender' and 'Tenure' auxiliary covariate CV_u 'during' points are at call one, due to CV_u non-significance (the 'after' points are again at the same call). The 'Ethnicity', 'HH structure' and 'Qualifications' points are mostly earlier than the 'Located in London/SE', 'Age' and 'Activity last week' points. Some are earlier than the numeric points, due to the 95% CI overlapping at CV_u differences larger than 0.02. The under-represented category Employed and 28 to 39 category CV_u points are later than the No qualifications, Asian and Single adult points, with differences from the numeric points due to the non-significance of the CV_u or the overlapping of CI at CV_u differences less than 0.02.

3.2. Evaluating CV based inference for survey variables with auxiliary covariate analogues

3.2.1. Census auxiliary covariate analogue category 'non-response biases'

We report estimated biases in Fig. 4, and their 95% CIs in Tables 5 & 6 in the on-line appendix. They are mostly consistent with the CVs. As expected, since the CVs predict (conditional) biases of the category means, overall (model A) and conditional (model B) (absolute) biases for the two-category auxiliary covariates are quantitatively similar to the covariate CV_us and CV_cs respectively. Correspondence is close for the 'Tenure', 'Located in London/SE' and 'Gender' overall biases: conditional biases can be slightly larger than the CV_cs. The 95% CI widths for bias tend to decline over calls, as with the CVs. Some significance differences exist: the 95% CIs for the 'Gender' call one overall bias and the later call 'Gender' and 'Tenure' conditional bias exclude zero.

Moreover, for the studied multi-category auxiliary covariate categories qualitative similarities at least exist between the (absolute) bias estimates and the category CVs (the CVs predict bias minima, with under-estimation less when contributions to the covariate inequalities are large). The Asian and No qualifications overall biases correspond with the CV_us, with conditional biases slightly larger and smaller than the CV_cs respectively. The Single adult HH, Employed and 28 to 39 biases are larger than the CVs: the last two differences decline over the calls because contributions to the covariate inequalities increase. The widths of the 95% CIs for bias also tend to decline, with similar significance for the CVs. In addition, biases are smaller than the relevant covariate CVs, which in these cases predict category bias maxima, and all biases are smaller than the overall CVs, which predict (survey wide) category bias maxima.

3.2.2. Census auxiliary covariate analogue category 'non-response bias' PC points

The estimated overall bias and CV_u PC points also mostly correspond (Table 3). With the numeric identification methods, the two-category auxiliary covariate bias and covariate CV_u points are at the same calls. The same occurs for multi-category auxiliary covariate categories, except for with Employed, for which the bias points are two calls later. With the inferential methods, the two-category auxiliary covariate bias and the covariate CV_u points are at the same calls except for the 'Tenure' 'during' point, which is at call two due to the significance of the call one estimate. For the multi-category auxiliary covariate categories, the points are at the same call, or the bias points are one to two calls earlier. Concerning the inferential points, though the statistical power issues associated with their identification are less problematic in our analyses (see section 2.1.4.), we note that while some of the bias points are earlier (the CIs are wider), correspondence between the CV_u and bias points is similar to that reported here when subsets of the dataset with 5000 and 10000 subjects are analysed (results not shown).

3.2.3. Survey variable – census auxiliary covariate analogue similarity

We report the category proportions for the survey respondents in the two data sources in Fig. 5. The values are as expected regarding the implied biases (the census sample values are mostly higher) and the changes over calls: they increase for categories becoming less under-represented and decrease for those becoming more so. They are also consistent with the survey variable – census auxiliary covariate analogue similarity. The Male, London/SE, Not owned HH and 28 to 39 proportions in the two sources are indistinguishable in the plots. Minor differences exist (mainly at early calls) for Single adult HH, Employed, No qualifications and Asian. For the first five categories, the Z tests for differences are all non-significant at the 0.05 level (see Table 7 in the on-line appendix). For the rest, the differences are significant after calls three to four: given the point estimates, as mentioned when identifying the PC points (see section 2.1.4), this is due to the increasing size of the respondent dataset.

4. Discussion

We evaluate the performance of the Coefficients of Variation of the response propensities (CVs) when monitoring the risks of survey variable non-response biases during survey data collection. CVs quantify dataset representativeness in terms of variation in sample response propensities estimated given a fully observed auxiliary attribute covariate set correlated with the survey variables: high representativeness implies low bias risk. Practitioners are interested in using CVs to monitor survey data collection, but little research exists on how well they predict observed biases. We extend work on CV predictions concerning biases and how they inform dataset improvements. Next, we use CVs to quantify (changes in) UK Labour Force Survey (LFS) dataset representativeness over data collection, utilising linked survey sample census responses as auxiliary covariates. Then, we evaluate CV inferences about survey variables with analogues estimating the same quantities among the auxiliary covariates.

Regarding bias prediction, overall CVs approximate the maximal absolute standardised bias of survey variable means when non-response correlates maximally with the auxiliary covariates (de Heij et al. 2015). We show that partial unconditional and conditional covariate CVs (CV_us and CV_cs respectively), which decompose overall CVs to measure (conditional) deviations in response with respect to auxiliary covariates, also predict similar absolute standardised ‘non-response biases’ of category means for two-category auxiliary covariates. For similar multi-category covariates, category (focal vs. others) bias maxima are predicted. Category CVs predict category bias minima, with less under-estimation when contributions to covariate inequalities are large. These predictions have not previously been reported, and potentially increase the utility of CVs when assessing survey datasets (and others with missing data, for example linked datasets; e.g. Moore et al. 2018c). If the survey variables and auxiliary covariate analogues measure the same quantities, partial CVs can be used to make inferences about survey variable biases. Our empirical work, which we discuss below, tests this contention in the LFS.

Concerning informing dataset improvements, CV_us and CV_cs also measure deviations with regard to covariates from respectively MCAR and MAR given the other auxiliary covariates (Schouten et al. 2012). With statistical inference possible, they hence identify targets for collection method modifications: under-represented categories with significant CV_us and CV_cs (i.e. independent impacts), although categories with impacts also correlated with other covariates, as indicated by significant CV_us but non-significant CV_cs, should be considered as well. We show that for similar reasons CVs can help select auxiliary covariates to use in post-collection bias adjustments. Such adjustments generally assume response is MAR given a set of auxiliary covariates. To select the auxiliary covariate sets, Särndal & Lundström (2010) use the Coefficient of Variation of the weights (larger is better). This is equivalent to the overall CV when the weights and response propensities are similarly estimated (for instance, by logistic regression) or the sample size is large, so significant CV_c identify covariates with independent impacts. Recognising this functionality also increases the utility of CVs when assessing dataset quality.

Our empirical work demonstrates the accuracy of CV based inference during data collection. We quantify LFS dataset representativeness by computing the overall CVs and auxiliary covariate CV_us and CV_cs after each attempt to interview non-respondents (the call record). We also identify phase capacity (PC) points after which further quality increases are limited and methods should be modified or data collection ended (e.g. Groves & Heeringa 2006). We consider stability of the CVs compared to previous call values (of use during collection to inform current sampling), and best values over the call record (of use after collection to inform future efforts). We use both numeric methods (do the CVs fall within a threshold of relevant values), and novel inferential methods (are the CVs non-significant or do the 95% CIs overlap) that we describe in section 2.1.4. Then, we evaluate CV based inference about the survey variables with auxiliary covariate analogues measuring the same quantities. First, we compare auxiliary covariate partial CVs to logistic regression based estimates of covariate category standardised 'non-response biases'. Second, we assess the survey variable – auxiliary covariate similarity by comparing the survey respondent category proportions given each data source. Pertinent to the performance of the CVs (we discuss other findings below), inference matches that from estimates of bias. The two-category auxiliary covariate CVs and estimated biases (and the PC points) correspond. The multi-category auxiliary covariate category CVs are smaller than the estimated biases (the PC points are mostly similar), and the covariate and overall CVs are larger. Moreover, the differences in the category proportions for the survey respondents between the data sources are slight, implying generalisability of inferences.

These findings indicate CVs are of utility as tools for monitoring survey data collection. Valid inference about the non-response biases of survey variables enables informed decisions about the methods to use to maximise final dataset quality. We hence recommend them to practitioners, and in Table 4 provide guidance on using them to monitor data collection in empirical scenarios in the form of a set of steps that should be included in analyses (see Schouten et al. 2012 for similar advice on assessing final dataset quality). Depending on the aims of monitoring, not all steps will be relevant. We note though that such aims are likely to depend on analysis findings: for example, without a PC point existing, practitioners may not have the resources to modify collection methods. We also note that if the aim is to modify methods to improve the dataset, after implementing modifications the CVs can be computed again to quantify their impact.

We do though make several comments about our evaluations and their implications. First, one limitation is that we do not evaluate CV based inferences about survey variables without auxiliary covariate analogues. Often, but not always (for example, the LFS is used to estimate UK employment rates; see ONS 2014), these are the main focus of a survey. We will be undertaking these evaluations in following research. Second, it should be noted that auxiliary covariate analogue partial CVs will perform best in predicting biases in survey variables when data sources do measure the same quantities. Dissimilarities may occur due to non-contemporary sources, or if the information requested or reported differs: the latter, caused by the LFS interviewers eliciting more accurate responses than the self-reported census, explains the slight differences found in our work between survey and census 'Ethnicity' Asian and No 'Qualifications' survey respondent category proportions (see Moore et al. 2018b). Hence, if possible survey variable – auxiliary covariate analogue similarity should be assessed before using CVs for this purpose.

Regarding our other findings, we study the LFS individual dataset, extending work on the household dataset (Moore et al. 2018a) to the sample unit. The overall CVs imply dataset non-representativeness decreases at a declining rate over calls, and is substantial when collection ends. The partial CVs suggest inequalities (biases) associated with six of the eight auxiliary covariates, with a range of under-represented categories (see section 3.1 for details and causes). Some impacts decline, others do not, and they are often independent. Regarding improving datasets, such categories are targets for data collection method modification (similarly, covariates should be used in post-collection adjustments). The identified PC points inform on when modifications should take place. The overall CV points are at calls four to five. The partial CV points, of use if separate targeting is possible, vary depending on category from calls one to six: similar variation in when estimate stabilise is found in other studies monitoring survey data collection (e.g. Petychev et al. 2009). The 'during' and 'after' rule points exhibit some differences, as do the points identified by the numeric and inferential methods. The latter have received little attention in the context of CVs: as found in research using other estimators (Lewis 2017), our work, utilising simple 95% CI based tests, suggests that selecting significance levels suitable over all respondent dataset sizes is an issue with their use in empirical scenarios (see also section 2.1.4).

We lack information on alternative collection methods, so cannot advise further on improvements to the LFS dataset. What is useful though is to utilise overall CV points to identify when to end current data collection, so resources can be otherwise invested to improve quality or make cost savings. The identified points are slightly earlier than the LFS household dataset points (see Moore et al. 2018a), in part due to us excluding subjects aged over 64 (who do not answer some survey items) from analyses, and represent reductions in calls made of 12-19%. Substantial savings are likely from such reductions. Similar CV based results are also found for other European and UK surveys (Lundquist & Särndal 2013; Correa et al. 2016; Moore et al. 2018a). Hence, we end by recommending that more attention is paid to whether the number of calls currently made to social survey non-respondents are needed to maintain dataset quality.

5. Acknowledgements

This research was funded by the ESRC National Centre for Research Methods, Workpackage 1 (grant reference number ES/L008351/1) and the ESRC Administrative Research Centre for England (ADRCE) (grant reference number ES/L007517/1). This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

6. References

Barbosa, D. (2014) *Using Multilevel Models to Investigate Interviewer Effects on Nonresponse Bias and Measurement Error*. PhD Thesis, University of Southampton, UK.

Beaumont, J. F., Bocci, C. and Haziza, D. (2014) An adaptive data collection procedure for call prioritization. *J. Off. Stat.*, 30, 607-621. DOI: 10.2478/jos-2014-0040

Bethlehem, J. G. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Correa, S., Durrant, G.B. and Smith, P.W.F. (2016) Assessing non-response bias using call record data with applications in a longitudinal study. *S3RI Technical paper, University of Southampton. Southampton, UK, 38pp.*

de Heij, V., Schouten, B. and Shlomo, N. (2015) RISQ Manual 2.1: Tools in SAS and R for the computation of R indicators and partial R indicators. Available from: www.risq-project.eu.

de Leeuw, E. & de Heer, W. (2002) Trends in household survey nonresponse: A longitudinal and international perspective. In: *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little). pp. 41-54. New York: Wiley.

Durrant, G. B. & Steele, F. (2009) Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *J. R. Statist. Soc. A*, 172, 361-381. DOI: 10.1111/j.1467-985X.2008.00565.x

Durrant, G.B., Groves, G., Staetsky, L. and Steele, F. (2010) Effects of interviewer attitudes and behaviours on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36. DOI: 10.1093/poq/nfp098

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, *J. R. Statist. Soc. A*, 174, 4, 1029-1049. DOI: 10.1111/j.1467-985X.2011.00715.x

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, *J. R. Statist. Soc. A*, 176, 251-269. DOI: 10.1111/j.1467-985X.2012.01073.x

Fox, J. (2003) Effect displays in R for generalised linear models. *Journal of Statistical Software*

15, 1–27.

Fox, J. and J. Hong (2009) Effect displays in R for multinomial and proportional-odds logit models? Extensions to the effects package. *Journal of Statistical Software* 32, 1–24.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R. M. and Heeringa, S. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. Roy. Stat. Soc. Ser. A.*, 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.

Hastie, T. J. (1992) In: *Statistical Models in S* (eds. J. M. Chambers and T. J. Hastie), Wadsworth.

Jackson, C. (2011) Multi state models for panel data: the 'msm' package for R. *Journal of Statistical Software* 32, 1–24.

Kreuter, F. (2013) Facing the non-response challenge. *Annals of the American Academy of Political and Social Science* 645: 23-35. DOI: 10.1177/0002716212456815

Lewis, T. (2017) Univariate tests for phase capacity: tools for identifying when to modify a survey's data collection protocol. *J. Off. Stats*, 33, 601-624. DOI: 10.1515/jos-2017-0029

Little, R. J. and Rubin, D. B. (2002) *Statistical inference with missing data.* Wiley, New York.

Little, R. J. A. and S. Vartivarian (2005) Does weighting for nonresponse increase the variance of survey means? *Surv. Meth.* 31: 161-168.

Lundquist, P. and Särndal, C-E. (2013) Aspects of responsive designs with applications to the Swedish Living Conditions Survey. *J. Off. Stats*, 29, 557-582. DOI: 10.2478/jos-2013-0040

Moore, J. C., Durrant, G. B. & Smith, P. W. F. (2018a) Dataset representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. *J. Roy. Stat. Soc. Ser. A.*, 181, 229-248. DOI: 10.1111/rssa.12256

Moore, J. C., Smith, P. W. F. and Durrant, G. B. (2018b) Discussion of paper by Hand, Statistical challenges of administrative and transaction data. *J. Roy. Stat. Soc. Ser. A.* 181, 584-585. DOI: 10.1111/rssa.12315.

Moore, J. C, Smith, P. W. F. and Durrant, G. B. (2018c) Business datasets and record linkage: Correlates of linkage and estimating risks of non-linkage biases. *J. Roy. Stat. Soc. Ser. A.*, 181, 1211-1230. DOI: 10.1111/rssa.12342

Nishimura, R., Wagner, J. and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *Int. Stat. Rev.*, 84, 43-62. DOI: 10.1111/insr.12100

Oehlert, G. W. (1992) A note on the delta method. *The American Statistician*, 46, 27-29.

Olson, K. (2006) Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, 70, 737–758. DOI: 10.1093/poq/nfl038

ONS (2011) LFS User Guide – Volume 1: Background and Methodology. Available at:

http://discover.ukdataservice.ac.uk/catalogue/?sn=6782&type=Data%20catalogue.

ONS (2014) 2011 Census Variables: Part 1. Available at: http://www.ons.gov.uk/census

Parry-Langdon, N. (2011) *Social Survey Non-Response Update*. ONS Technical Report. Available at: http://www.ons.gov.uk/ons/dcp171766_240879.pdf.

Peytchev, A., Baxter, R. K. and Carley-Baxter, L. R. (2009) Not all survey effort is equal: reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly* 73, 785-806.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. Lindblad, M. (2010) Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods,* 4, 21-29.

Rao, R. S., Glickman, M. E. and Glynn, R. J. (2008) Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statist. Med.*, 27, 2196-2213. DOI: 10.1002/sim.3063

Särndal, C. E. and Lundström, S. (2010) Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 131-144.

Särndal, C. E. (2011) The 2010 Morris Hansen Lecture. Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C-E. and Lundquist, P. (2014) Balancing the response and adjusting estimates for nonresponse bias: complementary activities. *Journal de la Société Française de Statistique*, 155, 28-50.

Schouten, B. and Shlomo, N. (2017) Selecting adaptive survey design strata with partial R-indicators. *International Statistical Review*, 85, 143-163.

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.

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. *Int. Statist. Rev.*, 80, 382-399. DOI: 10.1111/j.1751-5823.2012.00189.x

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., Lundquist, P. and Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *J. Roy. Stat. Soc. Ser. A*, 179, 727-748. DOI: 10.1111/rssa.12152

Steele, F. and Durrant, G. (2011) Alternative approaches to multilevel modelling of survey noncontact and refusal, *Int. Statist. Rev.*, 79, 70-91. DOI: 10.1111/j.1751-5823.2011.00133.x

Wagner, J. R. (2008) *Adaptive Survey Design to Reduce Nonresponse Bias.* PhD diss., University of Michigan, Michigan.

Wagner, J.R and Raghunathan, T. E. (2010) A new stopping rule for surveys. *Statist. Med.*, 29, 1014-1024. DOI: 10.1002/sim.3834

Wagner, J. (2012) A Comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76, 555–75.

Table 1: Dataset construction and content. ‘Linked to census’, ‘Face to face interview’ and ‘With call records’, ‘Under 65’ and ‘Without item NRs’ are the number of (remaining) individuals and HHs with such characteristics, the last being the size of analytical dataset. ‘Interviewed’, ‘Refusal’ and ‘Non-contact’ are numbers of outcomes in the call 20 dataset.

	HHs	Individuals
Eligible	26322	64187
Linked to Census	25524	59897
Face to face interview	20123	41668
With call records	17760	36611
Under 65	14720	28383
Without item NRs	11491	21150
Interviewed (response)		12394
Refusal		2947
Non-contact		5809

Table 2: Studied survey variable – census auxiliary attribute covariate analogue pairs, and their categorisations.

Variable / covariate	Categories
Two category:	
Gender	1) Male; 2) Female.
Tenure	1) Owned; 2) Not owned.
Located in London/SE	1) No; 2) Yes.
Multi-category:	
Age	1) 16 to 27; 2) 28 to 39; 3) 40 to 51; 4) 52 & over.
Qualifications	1) NQF 4+; 2) NQF 3; 3) Apprenticeship; 4) NQF 2; 5) <NQF 2; 6) Other; 7) None; 8) Not recorded (NR).
Activity last week	1) Employed; 2) Unemployed; 3) Economically inactive (EI): Student; 4) EI: Retired; 5) EI: Ill / impaired; 6) EI: At home / other.
Ethnicity	1) White; 2) Mixed; 3) Black; 4) Asian; 5) Chinese; 6) Other; 7) NR.
Household (HH) structure	1) Single adult; 2) Single adult with children; 3) Couple; 4) Couple with children; 5) Other.

Table 3: Numeric threshold and statistical inference identified ‘during’ and ‘after’ rule design phase capacity (PC) points for selected census auxiliary covariate categories based on partial CV_us and where comparable transformed bias model A estimates.

	Numeric				Inferential			
	During		After		During		After	
	CV_u	Bias	CV_u	Bias	CV_u	Bias	CV_u	Bias
Covariate (two cat.):								
‘Located in London/SE’ London/SE	4	4	4	4	4	4	3	3
‘Tenure’ Not owned	2	2	1	1	1	2	1	1
‘Gender’ Male	2	2	1	1	1	1	1	1
Covariate (multi-cat.):								
‘Age’	4		4		2		4	
‘Activity last week’	5		5		4		4	
‘Ethnicity’	2		2		2		1	
‘Qualifications’	3		2		2		1	
‘HH structure’	2		1		2		1	
Category								
‘Age’ 28 to 39	3	3	4	4	4	2	5	4
‘Activity Last Week’ Employed	3	5	4	6	5	4	5	4
‘Ethnicity’ Asian	2	2	1	1	2	2	3	1
‘Qualifications’ None	3	3	2	2	2	2	2	2
‘HH structure’ Single adult	2	2	1	1	2	2	1	1

Table 4: The steps in an analysis utilising CVs to monitor survey data collection. Depending on the aim of monitoring, not all will be relevant in a given scenario, though these aims will likely depend on analysis findings (see also 'Discussion').

1. Select a set of auxiliary covariates observed for all sample subjects. Covariates should be correlated with (explain as much variation as possible in) the survey variables. An issue is likely to be availability: common sources of covariates such as administrative records and population registers tend to be limited in scope. We also note that given the impact of covariate categorisation on partial CV predictions, for key auxiliary covariate inequalities and especially when making inferences about biases in survey variables with auxiliary covariate analogues, binary coding (focal category vs. others) should be utilised (see also section 2.1.2).

2. Given the auxiliary covariate set, compute overall and auxiliary covariate partial CVs over collection as in section 2.2.2. We study a call record, but methods can be adapted, for instance to assess whether offering non-respondents another response mode improves the dataset.

3. If aiming to identify whether collection can be ended early, select methods (during or after collection, numeric (threshold) or inferential (significance level)) from those outlined in section 2.1.4 and seek to identify the overall CV PC point. If a survey variable with an auxiliary covariate analogue is the focus, use the same methods and the relevant CV_u.

4. If aiming to identify auxiliary covariates to use in post collection bias adjustments, check covariate CV significance at the overall CV PC point (if one exists) or in the final dataset. Exclude from sets covariates with non-significant CV_us, and those with non-significant CV_cs unless correlated with other similar covariates (see also section 2.1.2).

5. If aiming to identify targets for method modifications, identify auxiliary covariate categories with major (addressing large impacts most reduces the overall CV) independent impacts, i.e. those with large, significant CV_us and CV_cs, notwithstanding correlations between those with non-significant CV_cs (see also section 2.1.2). CVs at the overall CV PC point (if it exists) should be used, though if categories are separately targetable it may be possible to implement modifications at (earlier) relevant covariate / category CV_u points.

Figure 1: LFS dataset cumulative response rate (RR) over the call record and similar dataset overall CVs. See Table 1 in the one-line appendix for the CV 95% CIs.

Figure 2: LFS dataset partial unconditional and conditional auxiliary covariate CVs over the call record. See Tables 1 and 2 in the on-line Appendix for the CV 95% CIs.

Figure 3: LFS dataset partial unconditional and conditional auxiliary covariate category CVs over the call record: a) Age; b) Qualifications; c) Activity Last Week; d) HH structure; e) Ethnicity; f) Tenure; g) Located in London/SE; and h) Gender. See Tables 3 and 4 in the on-line appendix for the CV 95% CIs.

Figure 4: Partial CVs and model (A = overall, B = conditional) estimated standardised 'non-response biases' for the auxiliary covariate categories: a) 'Located in London/SE' Yes; b) 'Tenure' Not owned; c) 'Gender' Male; d) 'Age' 28 to 39; e) 'Activity Last Week' Employed; f) 'Ethnicity' Asian; g) 'Qualifications' None; h) 'HH structure' Single adult. The first three covariates have two categories, so the covariate CVs are comparators, with (as CVs are constrained to be positive) the model bias estimate absolute values reported. The other covariates are multi-category, so category CVs are comparators. With these, the CV_c is constrained to be positive, so model B based bias estimate absolute values are reported. See Tables 5 & 6 in the on-line appendix for the bias estimate 95% CIs.

Figure 5: Survey variable (dashed lines) and census auxiliary covariate analogue (thick solid lines) category survey respondent proportions over the call record for: a) 'Gender' Male and 'Located in London/SE' Yes; b) 'Tenure' Not owned and 'Age' 28 to 39; c) 'Activity Last

Week’ Employed and ‘HH structure’ Single adult. We also present census auxiliary covariate sample category proportions (thin solid lines).

Fig. 1

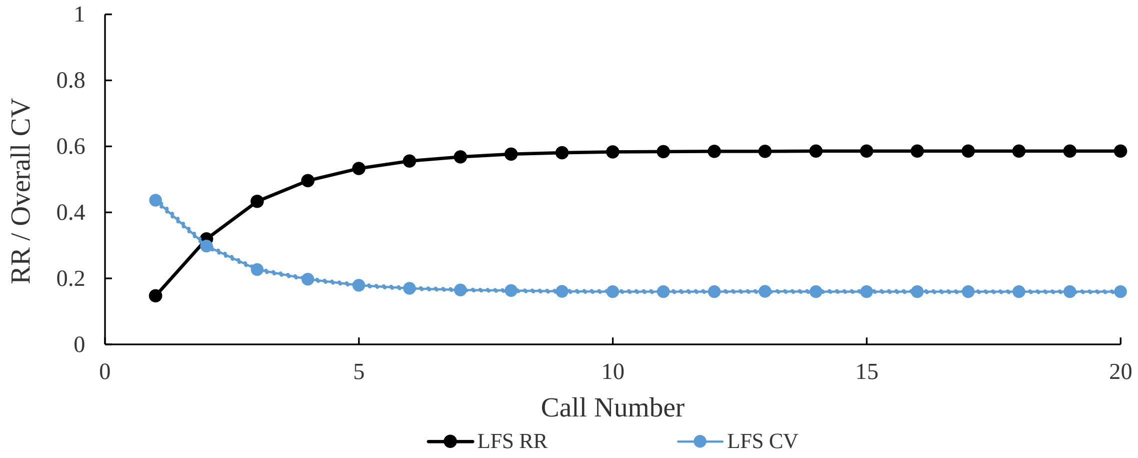
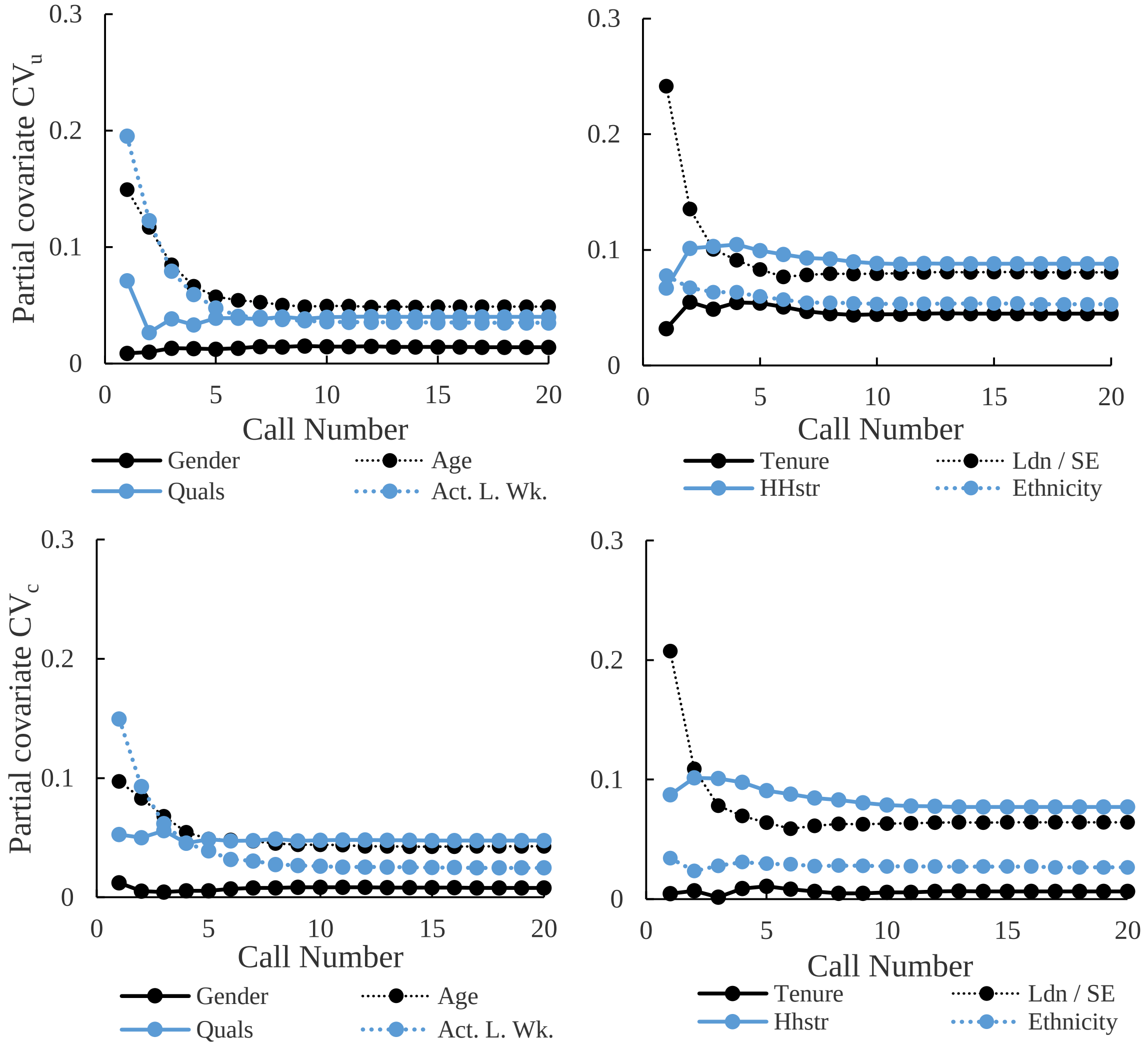


Fig 2



Partial covariate CV_u
0.3
0.2
0.1
0
0
5
10
15
20
Call Number
Gender
Age
Quals
Act. L. Wk.
0.3
0.2
0.1
0
0
5
10
15
20
Call Number
Tenure
Ldn / SE
HHstr
Ethnicity
Partial covariate CV_c
0.3
0.2
0.1
0
0
5
10
15
20
Call Number
Gender
Age
Quals
Act. L. Wk.
0.3
0.2
0.1
0
0
5
10
15
20
Call Number
Tenure
Ldn / SE
Hhstr
Ethnicity

Fig. 3

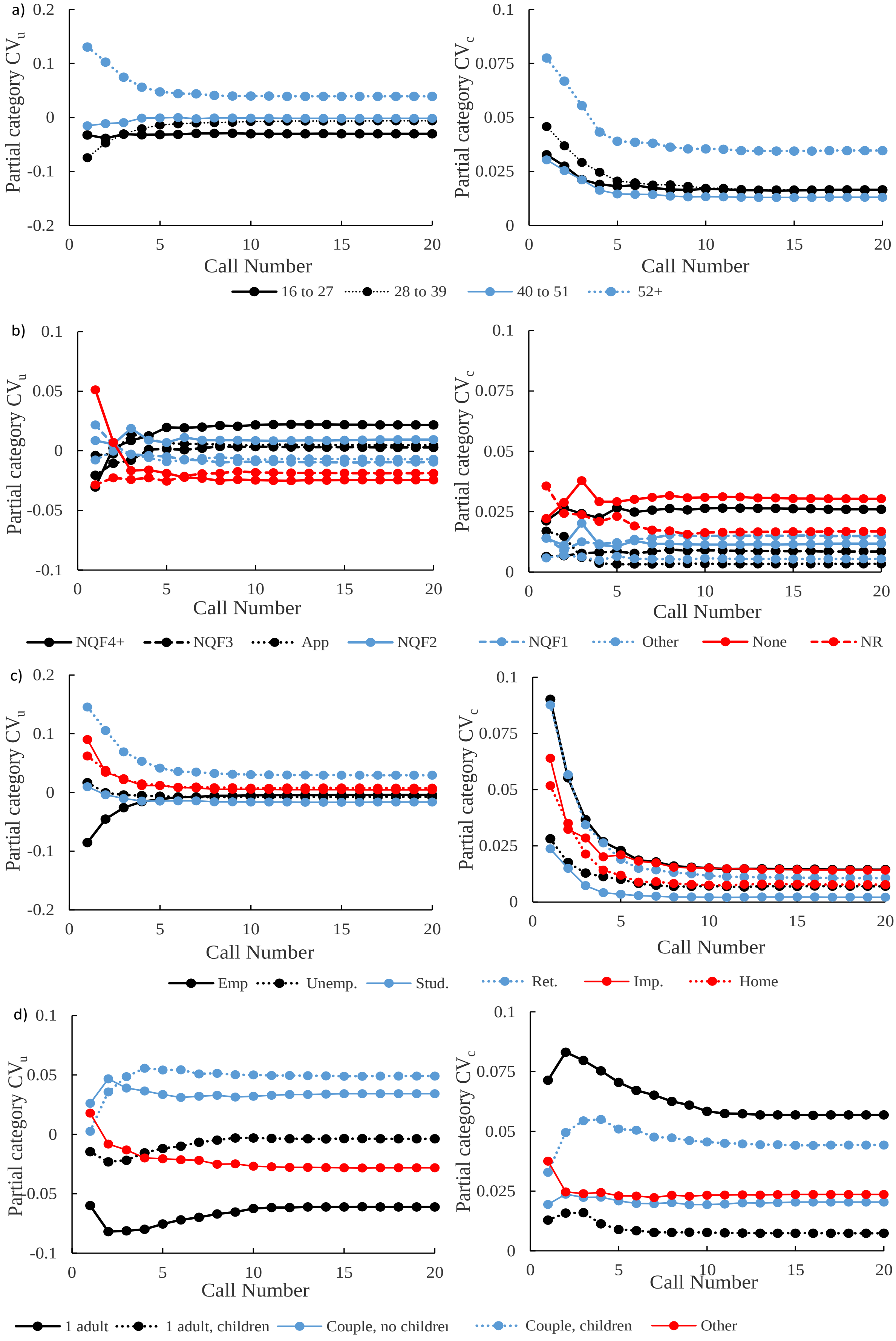
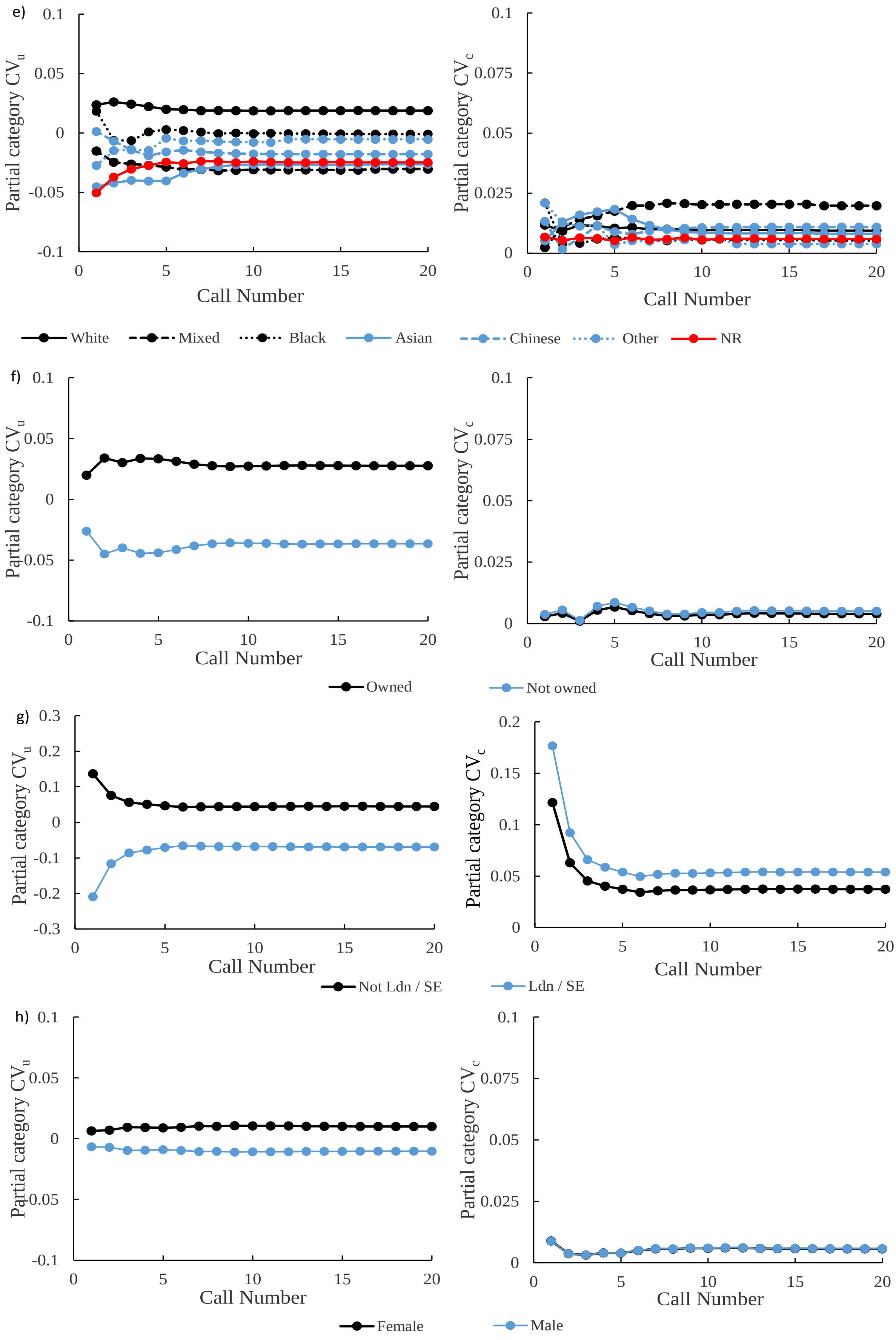
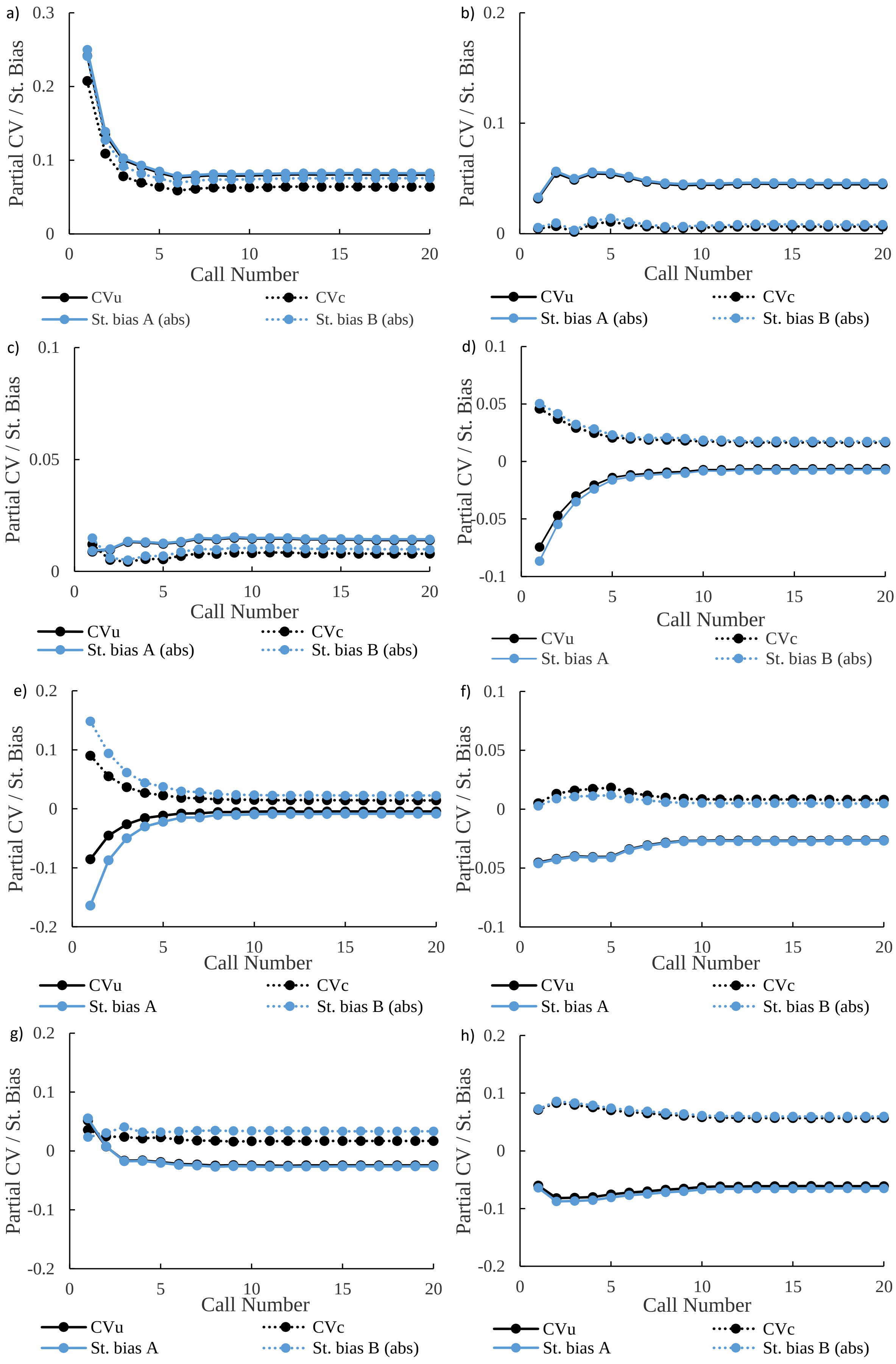

a)
Partial category CV$_u$
Partial category CV$_c$
Call Number
16 to 27
28 to 39
40 to 51
52+
b)
NQF4+
NQF3
App
NQF2
NQF1
Other
None
NR
c)
Emp
Unemp.
Stud.
Ret.
Imp.
Home
d)
1 adult
1 adult, children
Couple, no childrei
Couple, children
Other
0.2
0.1
0.075
0.05
0.025
0
-0.05
-0.1
-0.2
0
5
10
15
20

Fig 3 cont.



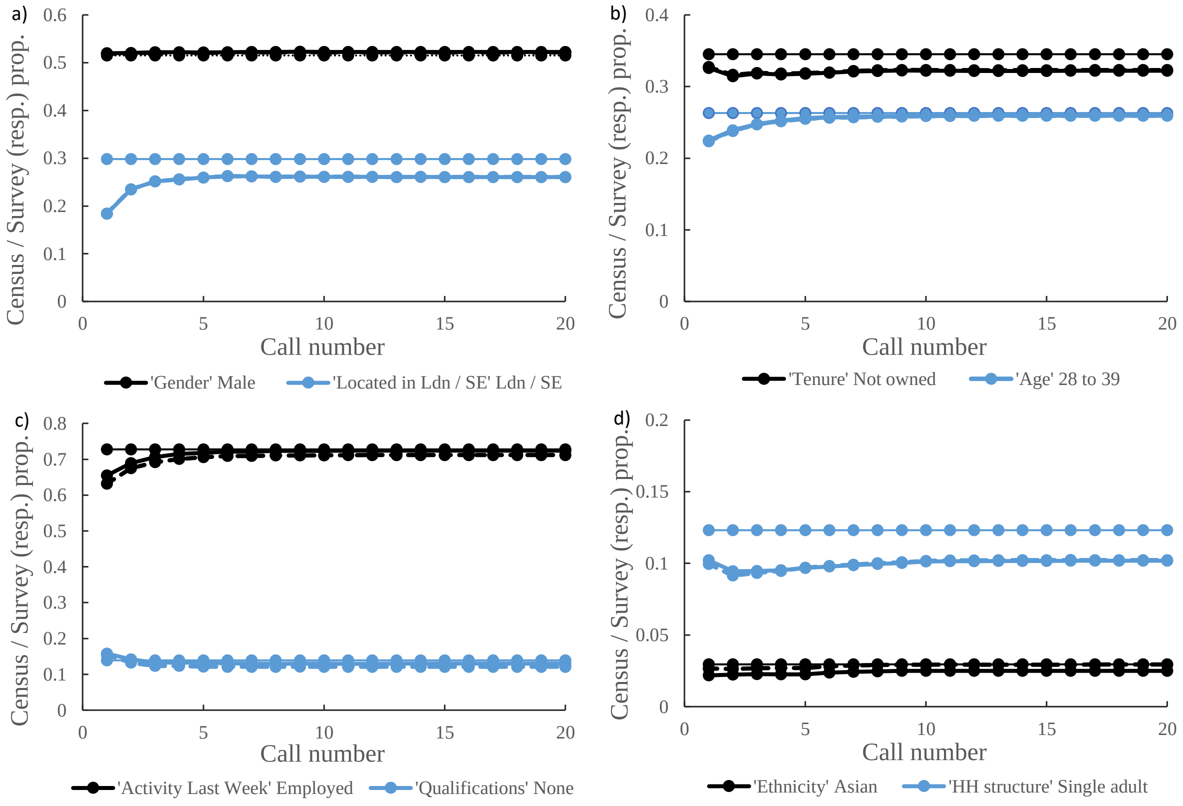
e)
0.1
0.05
0
-0.05
-0.1
Partial category CV_u
0
5
10
15
20
Call Number
0.1
0.075
0.05
0.025
0
Partial category CV_c
0
5
10
15
20
Call Number
White
Mixed
Black
Asian
Chinese
Other
NR
f)
0.1
0.05
0
-0.05
-0.1
Partial category CV_u
0
5
10
15
20
Call Number
0.1
0.075
0.05
0.025
0
Partial category CV_c
0
5
10
15
20
Call Number
Owned
Not owned
g)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Partial category CV_u
0
5
10
15
20
Call Number
0.2
0.15
0.1
0.05
0
Partial category CV_c
0
5
10
15
20
Call Number
Not Ldn / SE
Ldn / SE
h)
0.1
0.05
0
-0.05
-0.1
Partial category CV_u
0
5
10
15
20
Call Number
0.1
0.075
0.05
0.025
0
Partial category CV_c
0
5
10
15
20
Call Number
Female
Male

Fig. 4



a)
0.3
0.2
0.1
0
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A (abs)
St. bias B (abs)
b)
0.2
0.1
0
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A (abs)
St. bias B (abs)
c)
0.1
0.05
0
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A (abs)
St. bias B (abs)
d)
0.1
0.05
0
-0.05
-0.1
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A
St. bias B (abs)
e)
0.2
0.1
0
-0.1
-0.2
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A
St. bias B (abs)
f)
0.1
0.05
0
-0.05
-0.1
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A
St. bias B (abs)
g)
0.2
0.1
0
-0.1
-0.2
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A
St. bias B (abs)
h)
0.2
0.1
0
-0.1
-0.2
Partial CV / St. Bias
0
5
10
15
20
Call Number
CVu
CVc
St. bias A
St. bias B (abs)

Fig. 5



a)
0.6
0.5
0.4
0.3
0.2
0.1
0
Census / Survey (resp.) prop.
0
5
10
15
20
Call number
'Gender' Male
'Located in Ldn / SE' Ldn / SE
b)
0.4
0.3
0.2
0.1
0
Census / Survey (resp.) prop.
0
5
10
15
20
Call number
'Tenure' Not owned
'Age' 28 to 39
c)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Census / Survey (resp.) prop.
0
5
10
15
20
Call number
'Activity Last Week' Employed
'Qualifications' None
d)
0.2
0.15
0.1
0.05
0
Census / Survey (resp.) prop.
0
5
10
15
20
Call number
'Ethnicity' Asian
'HH structure' Single adult