ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)

Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice

Jamie C. Moore, Gabriele B. Durrant and Peter W. F. Smith

University of Southampton, UK

[Received September 2015. Final revision October 2016]

Summary. We consider the use of representativeness indicators to monitor risks of non-response bias during survey data collection. The analysis benefits from use of a unique data set linking call record paradata from three UK social surveys to census auxiliary attribute information on sample households. We investigate the utility of census information for this purpose and the performance of representativeness indicators (the *R*-indicator and the coefficient of variation of response propensities) in monitoring representativeness over call records. We also investigate the extent and effects of misspecification of auxiliary covariate sets used in indicator computation and design phase capacity points in call records beyond which survey data set improvements are minimal, and whether such points are generalizable across surveys. Given our findings, we then offer guidance to survey practitioners on the use of such methods and implications for optimizing data collection and efficiency savings.

Keywords: Adaptive and responsive survey designs; Coefficient of variation; Data collection efficiency savings; Phase capacity; *R*-indicators; Risk of non-response bias

1. Introduction

Survey methodologists no longer advocate maximizing response rates to minimize risks of non-response bias (see Olson (2006) and Kreuter (2013) for historic details). Rates have declined in the last 30 years (de Leeuw and de Heer, 2002) and have also been shown to be only weakly related to biases (Groves, 2006; Groves and Peytcheva, 2008). Instead, monitoring risks by quantifying variation in response between sample subgroups whose attributes are correlated with survey estimates is recommended, during data collection if paradata such as call records or details of other follow-up attempts are available. This can inform modifications to methods, to reduce such variation, and improve data set quality (by targeting underrepresented subgroups) and/or minimize costs (adaptive and responsive collection strategies, e.g. Groves and Heeringa (2006), Wagner (2008) and Peytchev *et al*. (2010)). Survey agencies are increasingly interested in employing this more refined approach to managing non-response bias risks, but reports of its use are still few, especially concerning monitoring during data collection, and available guidance is limited.

The above-described approach to managing non-response bias risks requires similarly motivated risk indicators (reviewed by Wagner (2012); see also Lundquist and Särndal (2013), Särndal and Lundquist (2014) and Correa *et al*. (2016)). One often-used type is representativeness indicators, which measure risks in terms of sample response propensity variation as

Address for correspondence: Jamie C. Moore, Administrative Data Research Centre for England and Department of Social Statistics and Demography, University of Southampton, Southampton, SO17 1BJ, UK.
E-mail: j.c.moore@soton.ac.uk

estimated by a statistical model given an auxiliary attribute covariate set. Low levels of variation imply representativeness and low risk of bias (see Schouten *et al.* (2016) for empirical support). Indicator computation requires auxiliary information on all sample units, concerning survey estimate correlates or sociodemographic attributes, which can be obtained from administrative data, a previous wave, census or population register. The most studied is the *R*-indicator, which is the transformed (0 to 1) standard deviation of response propensities, SD: $R = 1 - 2\,\mathrm{SD}$ (Schouten *et al.*, 2009, 2011, 2012). This form measures overall representativeness, enabling different surveys or waves to be compared given use of the same auxiliary covariate set. Partial decompositions measuring variation associated with factorial covariates also exist, enabling effects on representativeness to be assessed, for instance to identify target subgroups when modifying methods (see Schouten and Shlomo (2016)). Unconditional and conditional forms can be calculated, quantifying respectively the extent to which response with respect to a covariate is representative (a random sample) or conditionally representative (a random sample given stratifying covariates). Conditional variants thus enable detection of correlated effects, and so when modifying methods can ensure efficient targeting of (different) subgroups.

Guidance on several aspects of the use of these techniques to manage non-response bias risks is needed. To begin with, in previous reports sample information is from population registers, administrative data or previous waves (Lundquist and Särndal, 2013; Luiten and Schouten, 2013; Ouwehand and Schouten, 2013; Kappelhof, 2014; Correa *et al.*, 2016; Schouten *et al.*, 2016). In some countries including the UK, the first two sources of data do not exist, and the only available non-longitudinal information is from censuses. So far though, research on the use of such data is limited to identification of UK census-derived correlates of social survey non-response (Durrant and Steele, 2009; Steele and Durrant, 2011; Durrant *et al.*, 2010, 2011, 2013). Its utility for non-response bias risk monitoring is unknown.

There are also questions concerning representativeness indicator use when monitoring data collection. First, at low response rates possible response propensity variation is limited, so *R*-indicators may suggest that representativeness is highest early in call records (Schouten *et al.*, 2009). This can, for example, cause issues if identifying when to modify methods (see also below). An indicator with potentially better properties is the coefficient of variation (CV) of response propensities (Schouten *et al.*, 2009). The overall CV is SD divided by the mean propensity (low values imply representativeness), so it is less likely to be similarly affected by the response rate. It also provides a link to actual non-response biases, as it quantifies the maximal absolute standardized bias of a survey estimate mean when non-response correlates maximally to the utilized auxiliary covariate set. However, the CV is less studied than the *R*-indicator, especially partial decompositions (de Heij *et al.*, 2015), and comparisons of indicator behaviour over call records are rare (Lundquist and Särndal, 2013; Correa *et al.*, 2016).

Second, specifying auxiliary covariate sets for use over call records is problematic. Indicators are set specific, so the same set must be fitted at each call to isolate data set changes. Sets should include all available response propensity correlates: simulations suggest that exclusions lead to overall *R*-indicators comparatively overestimating representativeness (Shlomo *et al.*, 2012), and including non-correlates to underestimation and inflated indicator errors (Schouten *et al.* (2009); similar is expected with CVs). However, for a given sample size model selection methods should retain fewer covariates at low response rates, again because possible propensity variation is limited (Schouten *et al.*, 2009). Hence any set may be correctly specified (include only correlates) over only part of a call record, with sets correct at early call(s) likely to exclude later call data set correlates, and sets correct at later call(s) or specified without model selection likely to include (early call data set) non-correlates. Advice on covariate set specification given these considerations is lacking. We are unaware of any published empirical work on

propensity correlate changes over call records, or on the extent of set misspecification effects on indicators.

In addition, a focus when monitoring data collection is on identifying when continued use of current methods leads to minimal further increases (or even decreases) in the quality of data and modifications should be considered, termed reaching the design phase point capacity by Groves and Heeringa (2006) (see also Rao *et al*. (2008), Wagner and Raghunathan (2010) and Schouten *et al*. (2013)). However, reports of R-indicators and CVs discuss these phase capacity (PC) points only briefly, in the context of ending future data collection early given overall indicator stability compared with best values over (complete) call records (Correa *et al*., 2016). Points that are computed given partial indicators, and those computed given information only up to the current call (i.e. during collection), as necessary when historic data do not exist (e.g. Groves and Heeringa (2006)), are not presented, and so it is unknown how they compare. As well, whether PC points are generalizable from one survey to others, which is appealing to survey agencies given frequent legislative issues relating to linking sample information and also the costs of (realtime) monitoring, is unstudied.

We address these questions by using a unique data set linking details of attempts to interview households in three UK social survey samples to household attribute information from a concurrent census (a development of the Office for National Statistics 2011 Census Non-Response Link Study (CNRLS)). The data set enables monitoring of household level response (defined as at least one interview) during data collection, which we undertake by computing R-indicators and CVs at each call for each survey, considering 10 household attribute covariates in our analyses. First, we evaluate the utility of census data for this purpose. Second, we investigate auxiliary covariate retention in sets that are used in indicator response propensity estimation, by conducting logistic regression model selection given data sets after

(a) five interview attempts (early in data collection) and
(b) 20 attempts (the end of collection).

Third, we compare indicator behaviour and investigate auxiliary covariate set misspecification effects by computing indicators given sets (a) and (b) and also sets

(c) including all 10 covariates.

Fourth, we identify survey overall and partial CV stability-based PC points and evaluate their generalizability, both when entire call record information is available for their calculation (after collection) and when information exists only up to the current call (during collection). We then summarize our findings and offer guidance to survey practitioners on the issues considered.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Methods

2.1. Data sets

The CNRLS links January–July 2011 sample households from six UK social surveys to their March 27th, 2011, census records, providing attribute information whether they are interviewed or not (Parry-Langdon, 2011). We append call records, enabling monitoring during data collection, to three surveys:

(a) the Labour Force Survey (LFS), covering labour market topics (Office for National Statistics, 2011a),
(b) the Life Opportunities Survey (LOS), covering local facility use and leisure and employ-

Table 1. Data set construction and content†

	Results for the following surveys:		
	LFS	*LOS*	*OS*
Eligible households	27378	6896	6668
Linked to census	25524	6521	6260
Face-to-face interview	20514	6521	6260
With call records	18997	6469	6249
Interviewed (response)	12480	4533	3997
Refusal	1902	567	672
Non-contact	4615	1369	1580
Calls per household	8.67 (8.34)	8.32 (7.86)	9.33 (8.22)
Calls per successful interview	2.75 (1.95)	3.33 (2.26)	3.32 (2.32)

†'Linked to census', 'Face-to-face interview' and 'With call records' are the number of (remaining) households with such characteristics, the last being the analytical data set sizes. 'Interviewed', 'Refusal' and 'Non-contact' are numbers of outcomes in call 20 analytical data sets. We also present the number of calls that were made, as means and standard deviations (in parentheses) per household and per successful interview.

ment activity participation with a focus on the effects of impairment (Office for National Statistics, 2014a), and

(c) the Opinions Survey (OS), covering social and health topics (Office for National Statistics, 2011b).

The LFS and LOS randomly sample households and seek interviews with all household members. The OS randomly samples households within areas (postcode sectors) and seeks an interview with a single household member. Surveys are comparable both with respect to definitions of households and household level response (i.e. whether an(y) interview is obtained or not). The OS is a cross-sectional survey. The LFS and LOS are longitudinal, but to avoid sample attrition effects we consider wave 1 data only, so the data sets analysed are cross-sectional. Interviews are face to face in the LOS and OS, but in the LFS households can choose a telephone interview: a point that we return to below.

In the CNRLS, the Office for National Statistics link survey and census records by using automated and clerical household address matching. Linkage rates are high: 93.2% of households in the LFS, 94.5% in the LOS and 93.9% in the OS (Table 1). This means that we can study the majority of samples (although without non-linked household data we cannot completely rule out data set selection biases), using the rich suite of attribute covariates from the census (see Office for National Statistics (2014b)). Only households that were sampled close to the census date are included, so this information should reflect household attributes at the time of sampling. Hence, in this case census data are of great utility as a source of sample attribute information for monitoring response. We note *caveats* to this in other settings in Section 4.

We consider 10 auxiliary household attribute covariates in analyses (Table 2), chosen because analogues impact on 2001 CNRLS individual response propensities (see Durrant and Steele (2009)). 'Tenure', 'Accommodation type' and 'Cars available' are census household responses. 'HH economic status', 'HH structure', 'Ill health individual in HH', 'Impaired individual in HH', 'Retiree in HH' and 'English fluency in HH' are coded from individual census responses.

Table 2. Auxiliary household attribute covariates considered in the analyses, and categorizations

Covariate	*Categories*
HH economic status	1, all employed; 2, all unemployed; 3, all inactive; 4, mixed; 5, unknown
HH structure	1, 1 adult; 2, 1 adult, children; 3, couple, no children; 4, couple, children; 5, >2 adults, children or otherwise; 6, unknown
Accommodation type	1, house; 2, flat; 3, other; 4, unknown
Tenure	1, owned; 2, rented or other; 3, unknown
Cars available	1, none; 2, 1 car; 3, 2 cars; 3, 3 or more cars; 4, unknown
Ill health individual in HH	1, no; 2, yes
Retiree in HH	1, no; 2, yes
Located in London/SE	1, no; 2, yes
Impaired individual in HH	1, no; 2, yes
Anyone fluent in English in HH	1, yes; 2, no

'Located in London/SE' is a geographic identifier. The first five covariates are multicategory. 'Unknown' indicates no response. The others are binary, with no response coded as a negative.

The call record data detail outcomes of calls (non-contact, refusal or interview) to households (up to 20; Table 1). They do not exist for LFS telephone-interviewed households (approximately 20% of the sample), and some others (approximately 7% of the LFS sample; less than 1% in the LOS and OS; Table 1). After removing these households, the analysed LFS data set includes 18 997 households, the LOS 6469 households and the OS 6249 households. The final response rates were 65.7% in the LFS, 70.1% in the LOS and 64% in the OS. Analysis using the methods in the following sections suggests that households in houses, owner households, households with retirees and all inactive households are underrepresented in the analysed LFS data set compared with the all-linked households data set (results not shown), causing differences in covariate category household proportions compared with the OS and LOS data sets (see Table A1 in the on-line appendix). We consider how these impact on results in Section 3.4. We also note that in practical applications of the methods detailed here focusing on improving data sets, the effects of non-contact and refusal on representativeness must be quantified separately. The drivers of these two forms of non-response are likely to vary, as will their correlates. Hence, the effect of collection method changes on households (not) responding in each way will also probably differ (Durrant and Steele, 2009).

2.2. Representativeness indicators

Representativeness indicators quantify survey non-response bias risks in terms of sample response propensity variation. They are not directly related to (non-response biases in) specific estimates (Schouten *et al.*, 2012). Weighting can be applied to enable population level inference (see Roberts *et al.* (1987) for an introduction to the use of survey weights in propensity modelling), but here we study the linked sample (with call records). Some households are not linked to census data and are excluded from analyses, as are households without call records, so the weights supplied would not be useful. As well, ignoring sample design is justified because our interest is not in the population but in future data collection in the surveys (with the same designs: see Phipps and Toth (2012) for similar arguments in this context).

R-indicators are described by Schouten *et al.* (2009, 2011, 2012), and CVs by Schouten *et al.*

(2009) and de Heij *et al.* (2015). The overall R-indicator is the transformed (0–1) response propensity standard deviation SD: $R = 1 - 2\,\mathrm{SD}$, where

$$\mathrm{SD} = \sqrt{\left\{ \frac{1}{n-1} \sum_{i=1}^{n} (\hat{p}_i - \hat{\bar{p}})^2 \right\}},$$

n is the sample size, $\hat{p}_i$ the sample member i propensity and $\hat{\bar{p}}$ the mean propensity. Large indicators imply representativeness The overall CV is SD divided by $\hat{\bar{p}}$ and quantifies survey estimate mean maximum absolute standardized bias when non-response correlates maximally to the auxiliary covariate set x utilized (we emphasize that indicators are specific to this covariate set). Small values imply representativeness. Partial indicator decompositions allow propensity variation that is associated with auxiliary covariates and their categories to be quantified. Unconditional indicators measure univariate associations. The covariate of interest Z need not be in the covariate set x. The unconditional partial CV (we present CVs here: equivalent partial R-indicators are computed by removing the $\hat{\bar{p}}$ denominator terms) for covariate Z is

$$\widehat{\mathrm{CV}}_{\mathrm{u}}(Z, p_x) = \frac{\sqrt{\left\{ (1/n) \sum_{k=1}^{K} n_k (\hat{\bar{p}}_k - \hat{\bar{p}})^2 \right\}}}{\hat{\bar{p}}}, \tag{1}$$

where n_k is the size of covariate category k, and $\hat{\bar{p}}_k$ is the mean response propensity in k. Large values suggest substantial between-category propensity variability and non-representativeness that is associated with Z. The unconditional partial CV for category k of covariate Z is

$$\widehat{\mathrm{CV}}_{\mathrm{u}}(Z_k, p_x) = \frac{\sqrt{(n_k/n)}(\hat{\bar{p}}_k - \hat{\bar{p}})}{\hat{\bar{p}}}. \tag{2}$$

Indicators can be positive or negative, implying respectively overrepresentation or underrepresentation. The further they are from 0, the greater the effect. With conditional partial indicators, covariate Z must be in covariate set x. Indicators quantify non-representativeness associated with (the category of) Z conditional on other covariates, by comparing propensities given set x with and without Z. The conditional partial CV for covariate Z is

$$\widehat{\mathrm{CV}}_{\mathrm{c}}(Z, p_x) = \frac{\sqrt{\left\{ (1/n) \sum_{l=1}^{L} \sum_{i \in l} (p_i - \hat{\bar{p}}_l)^2 \right\}}}{\hat{\bar{p}}} \tag{3}$$

where $\hat{\bar{p}}_l$ is the mean response propensity of the lth of L cells resulting from cross-classification of x excluding Z and propensity modelling given this covariate subset. The conditional partial CV for category k of covariate Z is

$$\widehat{\mathrm{CV}}_{\mathrm{c}}(Z_k, p_x) = \frac{\sqrt{\left\{ (1/n) \sum_{l=1}^{L} \sum_{i \in l} h_i (p_i - \hat{\bar{p}}_l)^2 \right\}}}{\hat{\bar{p}}} \tag{4}$$

where h_i is an indicator detailing whether member i is in category k. In both cases, small indicators given large unconditional equivalents suggest effects also associated with other covariates. Large indicators imply uncorrelated effects.

Adjustments to overall and partial covariate indicators exist to account for sample-size-related biases caused by estimating propensities. Approximate R-indicator standard errors are also available, linearizing a variance estimator for SD derived by decomposing its distribution into

that due to sampling design and that due to propensity model parameter estimates (Shlomo *et al.*, 2012). For overall indicators, propensities are estimated given set x. For both partial indicators, they are estimated given a set including only Z. In addition, de Heij *et al.* (2014) derive overall CV standard errors, as the square root of the linearizing approximation:

$$\widehat{\text{var}}\{\text{CV}(p_x)\} \cong \frac{\text{SD}^2}{\hat{\bar{p}}^2}\left\{\frac{\widehat{\text{var}}(p)}{\hat{\bar{p}}^2} + \frac{\widehat{\text{var}}(\text{SD})}{\text{SD}^2} - 2\frac{\widehat{\text{cov}}(\hat{\bar{p}}, \text{SD})}{\hat{\bar{p}}\,\text{SD}}\right\} \tag{5}$$

where $\widehat{\text{var}}(p)$ is the estimated variance of the mean response propensity, $\widehat{\text{var}}(\text{SD})$ the estimated variance of the standard deviation of propensities and $\widehat{\text{cov}}(\hat{\bar{p}}, \text{SD})$ their estimated covariance. de Heij *et al.* (2014) assume that $\widehat{\text{var}}(p)$ is minimal and can be approximated by SD/n, that $\widehat{\text{var}}(\text{SD})$, renamed $\hat{S}^2$, can be approximated by the estimator that is derived by Shlomo *et al.* (2012) and that $\widehat{\text{cov}}(\hat{\bar{p}}, \text{SD})$ is negligible. Given this, they rewrite expression (5) as

$$\widehat{\text{var}}\{\widehat{\text{CV}}(p_x)\} \cong \frac{\text{SD}^2}{\hat{\bar{p}}^2}\left(\frac{\text{SD}^2}{n\hat{\bar{p}}^2} + \frac{\hat{S}^2}{\text{SD}^2}\right) = \frac{\hat{S}^2}{\hat{\bar{p}}^2} + \frac{\text{SD}^4}{n\hat{\bar{p}}^4}. \tag{6}$$

As with R-indicators, overall CV standard errors are computed with SD estimated given the whole auxiliary covariate set. We utilize this approach also to derive partial covariate CV standard errors, using the square root of the approximation (6) but for both unconditional and conditional indicators calculating SD given only Z, as with partial R-indicator errors. We extend the R code of de Heij *et al.* (2014) to produce partial CVs and these errors (as well as R-indicators, overall CVs and their errors). Our code is available on request. We note that de Heij *et al.* (2015) have recently similarly updated their code (a version in SAS is also available: see `www.risq-project.eu`). Their standard errors are derived by using a linearizing approximation from partial R-indicator errors. We present our errors here, as sometimes those of de Heij *et al.* (2015) are substantially inflated. This is because R-indicator errors are large when a covariate has minimal univariate effect on propensities and $\widehat{\text{var}}(p)$ is small, because of division of $\hat{S}^2$ by $\widehat{\text{var}}(p)$ in the derivation. Indictor point estimates are computed given a multivariate propensity model, so this can occur even if the covariate impacts non-trivially on representativeness (see Fig. A1 in the on-line appendix for errors of this type given our data sets). Beyond this, our errors are also around an order of magnitude smaller than those of de Heij *et al.* (2015) (results not shown).

2.3. Statistical analyses

We conduct two sets of statistical analyses. First, we investigate auxiliary covariate retention in sets for use in indicator response propensity estimation. We identify household attribute covariates impacting on response (a successful interview) propensities after

(a) five interview attempts (early in call records, when response rates are low) and
(b) 20 attempts (the end of data collection).

We use logistic regression to model propensities, fit main effects only, and retain only those covariates for which there is an increase in the Akaike information criterion (AIC) of more than 2 on removal from the final model (see Burnham and Anderson (2002)). Survey interviews may involve multiple calls: we consider the final call as the interview in these cases.

Second, we investigate representativeness indicator use to monitor data collection. For each survey, at each call we compute overall and partial R-indicators and CVs given auxiliary covariate sets (a) and (b) identified above, and also

(c) sets including all 10 covariates.

To study covariate set effects, we compare point estimates by calculating differences from no model selection set values, as percentages of the latter value (these sets are common comparators as they include all 10 covariates). This includes unconditional indicators for covariates that are not in the sets, but not conditional variants, which are calculable only for covariates in sets. We also compare overall indicator 95% confidence interval (CI) ranges, computing intervals as the indicator ± 1.96 times its standard error and calculating differences from no model selection set ranges as percentages of the latter value. We do not compare partial indicator 95% CI ranges as they are identical. As well, we consider statistical inference, studying whether overall indicator 95% CIs overlap and for partial variants also whether intervals span zero (implying (conditional) representativeness with respect to Z).

2.4. Phase capacity point identification

We identify stability-based overall CV PC points and partial unconditional CV PC points for covariates that are linked to substantial effects on overall data set representativeness. Inequalities underlying partial indicators are likely targets when modifying methods as their reduction will lead to the greatest increases in quality (Schouten *et al*., 2012; Schouten and Shlomo, 2016). We study information availability effects by using two identification rules:

(a) if CVs are within threshold a of best values over call records ('after' collection) and
(b) if CVs imply decreases in quality or are within a of the previous call value ('during').

We identify points when threshold a equals 0.01, 0.02 and 0.05. We also calculate the total calls that were made to samples saved by ending collection at overall CV points. We note that, when entire call record data exist, Schouten *et al*. (2013) present a framework for optimizing collection given alternative methods and quality–cost trade-offs, using representativeness indicators as quality measures. Points that were similar to our 'after' PC points, but also incorporating cost considerations, can be identified by treating them as possible alternative methods. However, such an analysis is beyond the scope of this paper: for a full representation, information on call costs as well as numbers is needed, which we lack.

3. Results

3.1. Response rate development

Survey household response rates increase similarly over call records, at decreasing rates with minimal increases after calls 9–11 (Fig. 1). The LFS call 1 response rate is higher but later increases smaller than in the OS and LOS (which has the highest final response rate).

3.2. Auxiliary covariate retention at different calls

In Table 3 we detail AIC-based model selection to identify household attribute covariates correlated with response propensity in the data sets after five and 20 interview attempts (the end of data collection; we present final model parameter estimates in Table A2 in the on-line appendix). All 10 covariates are never retained in covariate sets. Covariates retained differ both between call 5 and call 20 data sets and between surveys. Concerning the hypothesis that fewer covariates are retained at low response rates, as expected in the LFS and LOS fewer covariates are retained in call 5 sets. However, in the OS the reverse occurs, and some covariates are also retained only in the call 5 set in the LOS. Hence, the hypothesis is not always supported empirically.

3.3. Representativeness indicators and auxiliary covariate set effects

3.3.1. Overall indicators

In Fig. 1 we present survey overall R-indicators and CVs over call records given no model

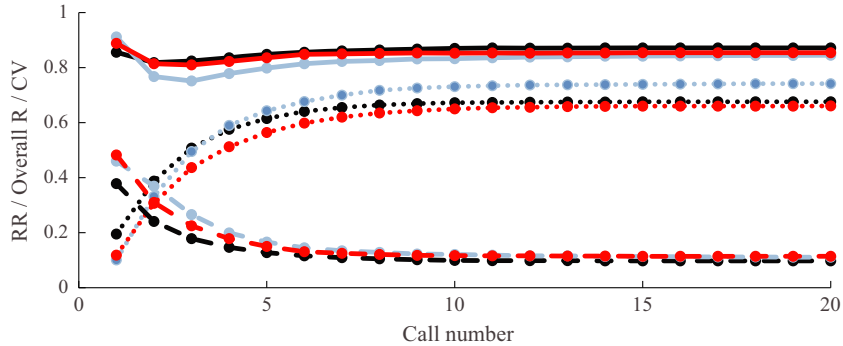

Fig. 1. Cumulative response rates, and overall *R*-indicators and CVs over call records in the three surveys given auxiliary covariate sets including all 10 household attribute covariates: • • ● • , LFS response rates; • • ● • , LOS response rates; • • ● • , OS response rates; ●—●, LFS *R*; ●—●, LOS *R*; ●—●, OS *R*; ●– –●, LFS CV; ●– –●, LOS CV; ●– –●, OS CV

Table 3. Covariates retained in logistic regression models of response propensity following AIC-based model selection on call 5 and call 20 data sets in each survey†

Covariate	*AICs for LFS*		*AICs for LOS*		*AICs for OS*	
	Call 5	*Final*	*Call 5*	*Final*	*Call 5*	*Final*
HH economic status	*25195*	*24110*	*8408.5*	*7736.4*	*8480.8*	*8070.8*
HH structure	*25348*	*24220*	*8409.8*	*7742.0*	*8472.8*	*8060.0*
Accommodation type	*25192*	*24117*	*8408.2*	7731.7	*8462.0*	8044.2
Tenure type	25176	*24112*	*8407.7*	*7736.0*	8461.4	8048.5
Cars available	25176	*24109*	8399.4	*7733.3*	*8472.2*	*8060.7*
Located in London/SE	*25236*	*24180*	*8421.3*	*7762.7*	*8474.8*	*8097.2*
English fluency in HH	25177	24109	8399.1	7730.3	*8461.8*	*8048.4*
Impaired individual in HH	25178	24109	8400.2	*7734.3*	8459.8	*8054.4*
Ill health individual in HH	25175	*24114*	8399.2	7732.3	*8465.1*	8045.8
Retiree in HH	*25199*	*24116*	*8426.9*	*7746.5*	*8465.4*	8047.3
Final model AIC	25176	24107	8399.2	7731.1	8459.0	8045.7

†AICs in italics indicate covariates that were retained in the models; AICs in normal text indicate covariates that were not retained.

selection auxiliary covariate sets including all 10 covariates. Indicators given the covariate sets identified in Section 3.2 are similar (CVs and 95% CIs are given as differences from no model selection set values in Table 4). *R*-indicators are initially large, implying high representativeness, decrease to call 3 and then increase at decreasing rates over the remaining calls (we term this the indicator trajectory). CVs decrease, implying increased representativeness, at decreasing rates over call records. Such *R*-indicator trajectories (equivalents are seen with partial variants; see Fig. A1 in the on-line appendix) can arise because possible propensity variation is limited at low response rates, which is an issue when modifying methods (see Section 1). That CVs, which are less likely to be similarly affected by the response rate and also quantifying maximum survey estimate mean absolute standardized bias when non-response correlates maximally to the utilized auxiliary covariate set, describe different changes, suggests that this is so here. Hence, hereafter we report only these indicators.

CVs are slightly lower in the LFS than in the other surveys and initially decrease less in the LOS than in the OS. 95% CI ranges are small (from about 0.002 to about 0.02). CV differences given

Table 4. Percentage differences between calls 5 ('CV5') and 20 ('CV20') auxiliary covariate set overall CVs and no model selection set values ('CV') in each survey†

Call	*Results for LFS*							*Results for LOS*							*Results for OS*						
	CV	*CV5*‡	*CV20*§	*CI−*	*CI+*	*CI5*	*CI20*	*CV*	*CV5*‡	*CV20*§	*CI−*	*CI+*	*CI5*	*CI20*	*CV*	*CV5*‡	*CV20*§	*CI−*	*CI+*	*CI5*	*CI20*
1	0.378	−1.60	−0.18	0.373	0.383	−3.23	−0.55	0.460	*−3.02*	1.48	0.450	0.471	−8.56	−2.11	0.483	0.37	−4.04	0.471	0.494	−1.06	−9.81
2	0.241	−0.14	−0.05	0.237	0.244	−1.08	−0.28	0.369	0.29	0.06	0.360	0.379	−0.96	−0.72	0.310	*−0.11*	*−5.57*	0.302	0.318	−0.89	−7.94
3	0.178	−0.01	0.06	0.176	0.181	−1.04	−0.18	0.266	−0.10	0.08	0.259	0.273	−1.01	−0.58	0.225	*0.06*	*−8.29*	0.219	0.231	−0.76	−9.24
4	0.147	−0.36	0.01	0.145	0.149	−1.49	−0.25	0.199	−0.27	−1.35	0.194	0.204	−1.26	−1.94	0.179	*−0.30*	*−6.21*	0.174	0.183	−1.16	−7.30
5	0.127	−0.90	0.06	0.125	0.129	−2.12	−0.25	0.166	−0.56	−1.80	0.162	0.170	−1.66	−2.53	0.150	*0.01*	*−5.22*	0.146	0.154	−1.02	−6.66
6	0.116	−1.99	0.07	0.114	0.117	−3.18	−0.26	0.145	−1.83	−1.79	0.142	0.149	−2.90	−2.85	0.131	−0.16	−3.34	0.127	0.134	−1.28	−5.44
7	0.110	−2.14	0.17	0.108	0.111	−3.38	−0.19	0.135	−1.83	−1.93	0.131	0.138	−3.08	−3.15	0.125	−0.44	−3.37	0.122	0.129	−1.48	−5.52
8	0.105	−2.17	0.17	0.103	0.106	−3.50	−0.21	0.128	−2.23	−1.79	0.125	0.131	−3.49	−3.19	0.121	−0.19	−2.53	0.118	0.124	−1.31	−4.95
9	0.102	−2.29	0.14	0.101	0.103	−3.66	−0.25	0.123	−2.07	−1.14	0.120	0.126	−3.47	−2.69	0.118	−0.59	−2.67	0.115	0.121	−1.62	−5.13
10	0.100	−2.39	0.13	0.098	0.101	−3.80	−0.27	0.121	−1.86	−1.22	0.118	0.124	−3.32	−2.80	0.117	−0.49	−2.19	0.114	0.120	−1.52	−4.75
11	0.098	−2.37	0.18	0.097	0.100	−3.82	−0.23	0.119	−2.05	−1.00	0.116	0.122	−3.52	−2.61	0.116	−0.62	−2.08	0.113	0.119	−1.63	−4.69
12	0.099	−2.41	0.18	0.097	0.100	−3.84	−0.23	0.116	−2.21	−1.22	0.113	0.119	−3.73	−2.87	0.116	−0.57	−1.85	0.113	0.119	−1.59	−4.51
13	0.098	−2.40	0.17	0.097	0.100	−3.83	−0.24	0.116	−2.19	−1.17	0.113	0.119	−3.74	−2.83	0.115	−0.66	−1.35	0.112	0.118	−1.66	−4.11
14	0.098	−2.35	0.17	0.097	0.099	−3.80	−0.24	0.113	−2.36	−1.12	0.111	0.116	−3.96	−2.85	0.115	−0.58	−1.20	0.112	0.118	−1.60	−3.98
15	0.098	−2.38	0.17	0.097	0.099	−3.83	−0.24	0.113	−2.30	−1.09	0.111	0.116	−3.92	−2.83	0.114	−0.64	−1.25	0.111	0.117	−1.65	−4.05
16	0.098	−2.42	0.17	0.096	0.099	−3.87	−0.25	0.112	−2.40	−1.10	0.110	0.115	−4.04	−2.85	0.114	−0.64	−1.25	0.111	0.117	−1.65	−4.05
17	0.098	−2.38	0.21	0.096	0.099	−3.84	−0.21	0.112	−2.38	−1.21	0.109	0.114	−4.04	−2.98	0.114	−0.64	−1.25	0.111	0.117	−1.65	−4.05
18	0.098	−2.38	0.21	0.096	0.099	−3.84	−0.21	0.112	−2.39	−1.22	0.109	0.114	−4.05	−3.00	0.114	−0.66	−1.27	0.111	0.117	−1.67	−4.05
19	0.098	−2.38	0.21	0.096	0.099	−3.84	−0.21	0.111	−2.55	−1.20	0.108	0.114	−4.20	−3.00	0.114	−0.66	−1.27	0.111	0.117	−1.67	−4.05
20	0.098	−2.38	0.21	0.096	0.099	−3.84	−0.21	0.111	−2.62	−1.23	0.108	0.113	−4.26	−3.03	0.114	−0.56	−1.35	0.111	0.117	−1.59	−4.12

†We also present similar indicator 95% CI range differences ('CI−' and 'CI+' detail the no model selection set range, 'CI5' and 'CI20' calls 5 and 20 set differences).
‡Values in italics indicate that calls 5 and 20 set 95% CIs do not overlap.
§Values in italics indicate that no model selection and call 20 set 95% CIs do not overlap.

different covariate sets reach approximately 10% in the OS but are mainly less than 4%, with CVs mostly smaller for sets with more covariates (Table 4). To investigate set misspecification effects, we compare indicators given different sets at calls 5 and 20, since we identify correctly specified sets including only propensity correlates at these calls in Section 3.2 and Table 3. An issue is that misspecified sets often both exclude correlates and include non-correlates. Concerning effects of excluding correlates, comparative overestimation of representativeness is predicted. One comparison exists where non-correlates are not also included, in the LFS at call 20. The CV given the (correlates excluded) call 5 set is smaller than that given the call 20 set, as expected.

Including propensity non-correlates in sets should lead to comparative underestimation of representativeness and inflated indicator errors. Comparisons where correlates are not also excluded involve no model selection set indicators at calls 5 and 20 and LFS call 20 set indicators at call 5. As expected, CVs and 95% CIs given these sets are larger than those given correct sets, except with LFS no model selection set CVs at call 5. Differences tend to be smaller than when correlates are excluded. Hence, covariate set misspecification effects are mostly, but not always, as hypothesized. Regarding statistical inference, small CV differences given different sets mean that their 95% CIs rarely fail to overlap (Table 4).

3.3.2. Partial indicators

Overall CV decompositions suggest similar effects on representativeness associated with household attribute covariates in each survey. We describe these by using covariate and selected covariate category partial CVs given no model selection covariate sets (Figs 2 and 3), though we also mention covariate indicators given the other sets identified (presented as differences from no model selection set values in Tables A3–A8 in the on-line appendix). 'Ill health individual in HH', 'Impaired individual in HH' and especially 'Retiree in HH' and 'HH economic status' partial unconditional CVs (CV_us) are initially high, implying substantial univariate associations with response propensity variation, and then decrease at decreasing rates over call records. 'Located in London/SE' CV_us are similar, though they reach minima and then increase slightly in the OS. 'HH structure' CV_us in the LOS and OS are also similar, but in the LFS first increase slightly and then decrease over call records. 'Accommodation type' CV_us decrease slightly, after first increasing in the LOS and OS. 'Cars available' CV_us decrease, from a high initial value in the LOS, increase and then decrease slightly again. 'Tenure' CV_us first increase (less so in the LFS) and then decrease slightly. 'English fluency in HH' CV_us are minimal.

Covariate category CV_us suggest 'Ill health individual in HH', 'Impaired individual in HH', 'Retiree in HH' and 'HH economic status' impacts arise because households that are categorized as no in the first three cases and all employed in the last are initially underrepresented in data sets (later indicator decreases imply that many of these are interviewed eventually). Partial conditional CVs (CV_cs) for these covariates (categories) are mostly much smaller than CV_us, suggesting that impacts are correlated (the exceptions are comparable OS 'HH economic status' CV_us and CV_cs given the call 20 covariate set, which may be due to its excluding 'Retiree in HH'). Named categories do to an extent identify overlapping sample subgroups (for instance, retirees are unlikely to be employed), probably differing in how contactable (any) household members are. 'Accommodation type', 'Cars available' and 'Tenure' impacts, due respectively to flats, multicar and non-owner households being underrepresented (not shown) possibly reflect such differences also, with CV_cs also smaller than CV_us. In addition to this (single) impact on representativeness, two covariates have large CV_us and CV_cs, implying impacts that are not linked to other covariates. Households that are 'Located in London/SE' are underrepresented. 'HH structure' impacts are due to single-LFS-adult households being underrepresented and

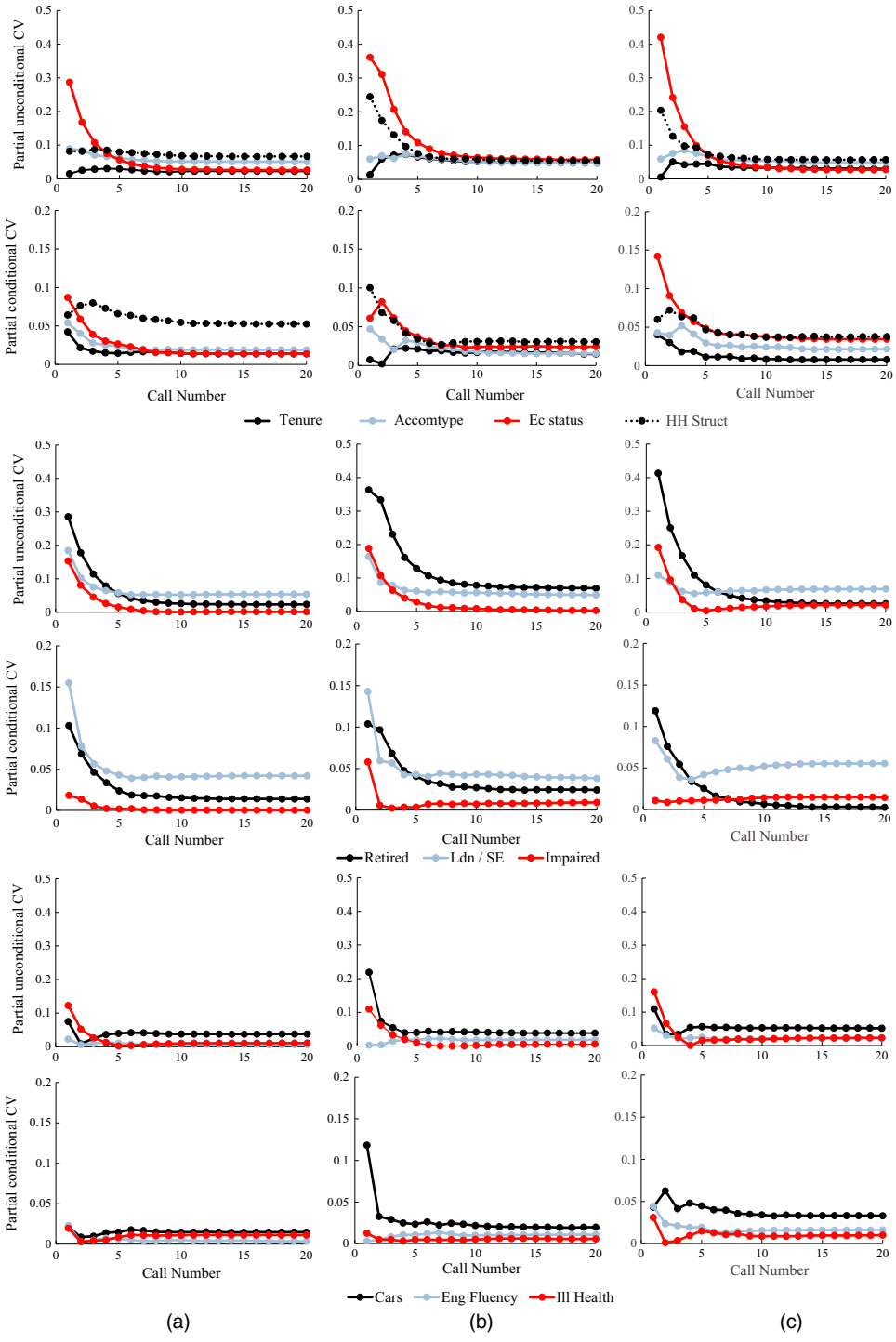

Fig. 2. Unconditional and conditional partial by covariate CVs over call records in the three surveys for household attribute covariates given auxiliary covariate sets including all 10 household attribute covariates (each column of graphs details indicators for a survey): (a) LFS; (b) LOS; (c) OS

LOS and OS couple, no-children households being overrepresented. Both these impacts are substantial at the end of data collection.

Concerning covariate partial CVs given different covariate sets, if the covariate is in both sets CV_us differ slightly (less than 2.5%) in ways that are identical to other similar covariates (see Tables A3–A8 in the on-line appendix). These differences reflect (differential) indicator bias adjustment, as equivalent unadjusted values differ negligibly (results not shown). If the covariate is not in both sets, CV_u differences are often greater than 50%, and once in the LOS approximately 14000%, with signs varying between covariates and over call records. CV_cs, calculable only for set members, always differ, mostly by less than 50%. Indicators are mostly minimal when differences are large though (about 0.000 01 in the LOS example), and indeed all actual differences are mainly small (reasons for OS 'HH economic status' CVs are given earlier). We again use call 5 and 20 indicators to study set misspecification effects. If a response propensity correlate is excluded, its CV_u is mostly, but not always, comparatively underestimated, but if a non-correlate is included effects on its CV_u vary (relevant comparisons are identifiable in Table 3). CV_us for other covariates (correlates) in sets given such exclusions or inclusions differ because of bias adjustment only, as noted above, but effects on CV_us for those (non-correlates) that are not in sets vary (based on the smaller relevant comparison set described in 'Overall indicators'). On the basis also of this smaller comparison set, set member CV_cs are mostly, but not always, overestimated when correlates are excluded, and underestimated when non-correlates are included, because of greater conditioning with larger sets. 95% CI ranges are small (from about 0.001 to 0.01). Regarding statistical inference, this means that indicator 95% CIs given different sets often do not (never with CV_cs) overlap. CV_u 95% CIs rarely span zero, at times doing so given one set but not others. CV_c 95% CIs never span zero.

3.4. Phase capacity points

We present indicator-stability-based overall and selected partial unconditional covariate CV PC points given 'after' and 'during' data collection identification rules and various rule thresholds a in Table 5. We illustrate results by using points when $a = 0.02$. Overall CV 'after' rule points are later in call records than 'during' rule points, and LOS points later than LFS and OS points, which are similar. Ending collection at these points saves the greatest percentage of the total calls made in the LOS (also see Table 5). Call savings range from 7% to 18%.

Our earlier analyses suggest three substantial effects on data set representativeness (see Section 3.3). We present unconditional partial CV PC points for the covariates 'Located in London/SE' and 'HH structure', which are linked to separate effects, and 'HH economic status' and 'Retiree in HH', which are linked to the same effect and so should have similar points. Points mostly differ from overall CV points and from each other, being earlier for the first two covariates because CVs decrease minimally over or are near minima early in the call record (points for the last two covariates are similar, as expected). Points tend to be later given 'after' than 'during' identification rules, but exceptions include OS 'Located in London/SE' and LFS 'HH structure' (though the latter is due to a previous call value being needed with 'during' rules). Some variability exists between covariates, but points are later in the LOS than in the LFS and OS, similarly to overall CV points. Both overall and partial CV points exhibit similar patterns when thresholds a equal 0.01 or 0.05. Points are earlier, and more calls are saved given overall CV points, as a is increased. A qualifier to our survey comparison results is that some differences exist between analysed LFS sample attribute category proportions and those in LOS and OS samples (see Section 2.1). However, it is LOS PC points that differ from the others: LFS points should do so if sample composition differences are important.

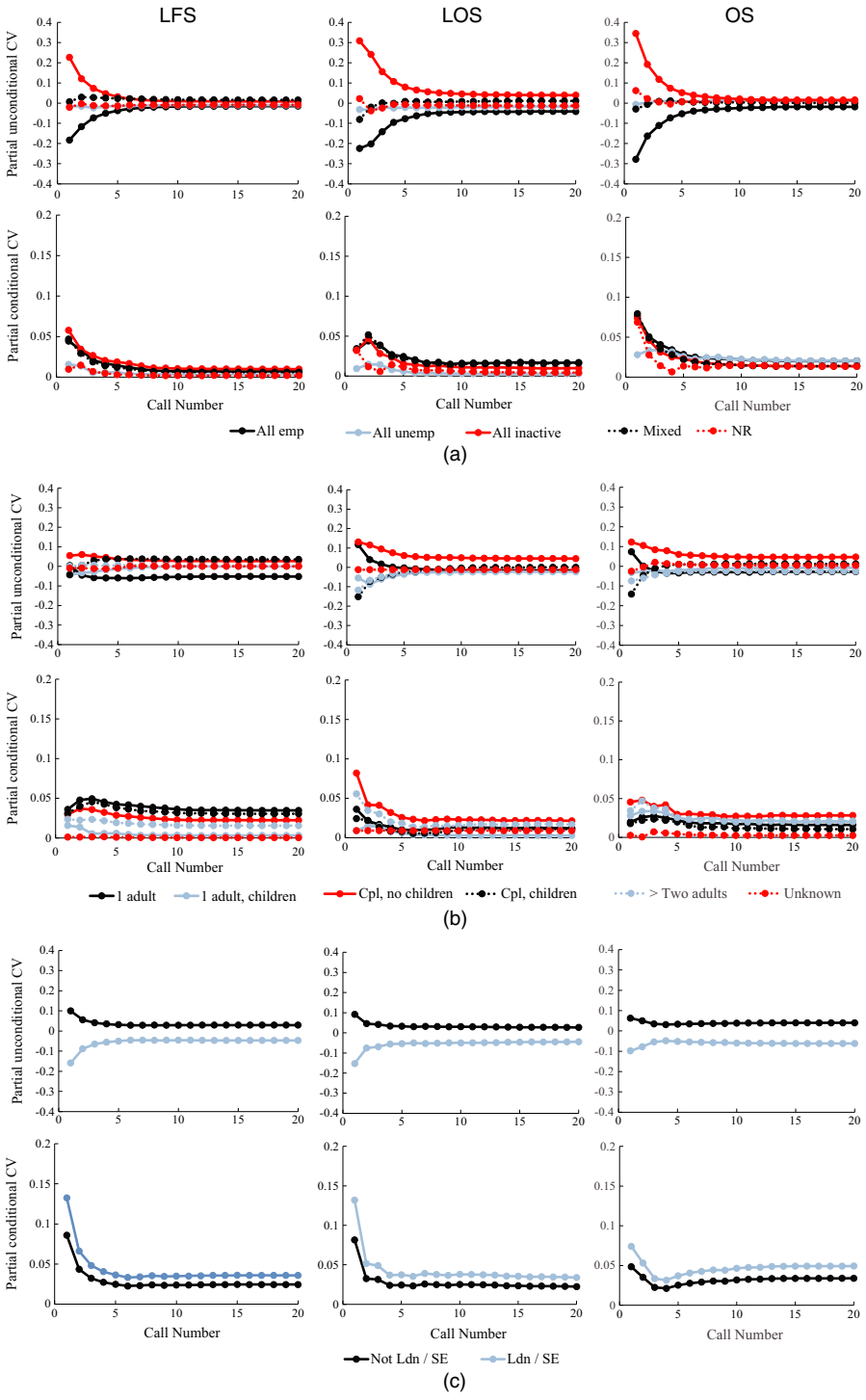

Fig. 3 (*continued*)

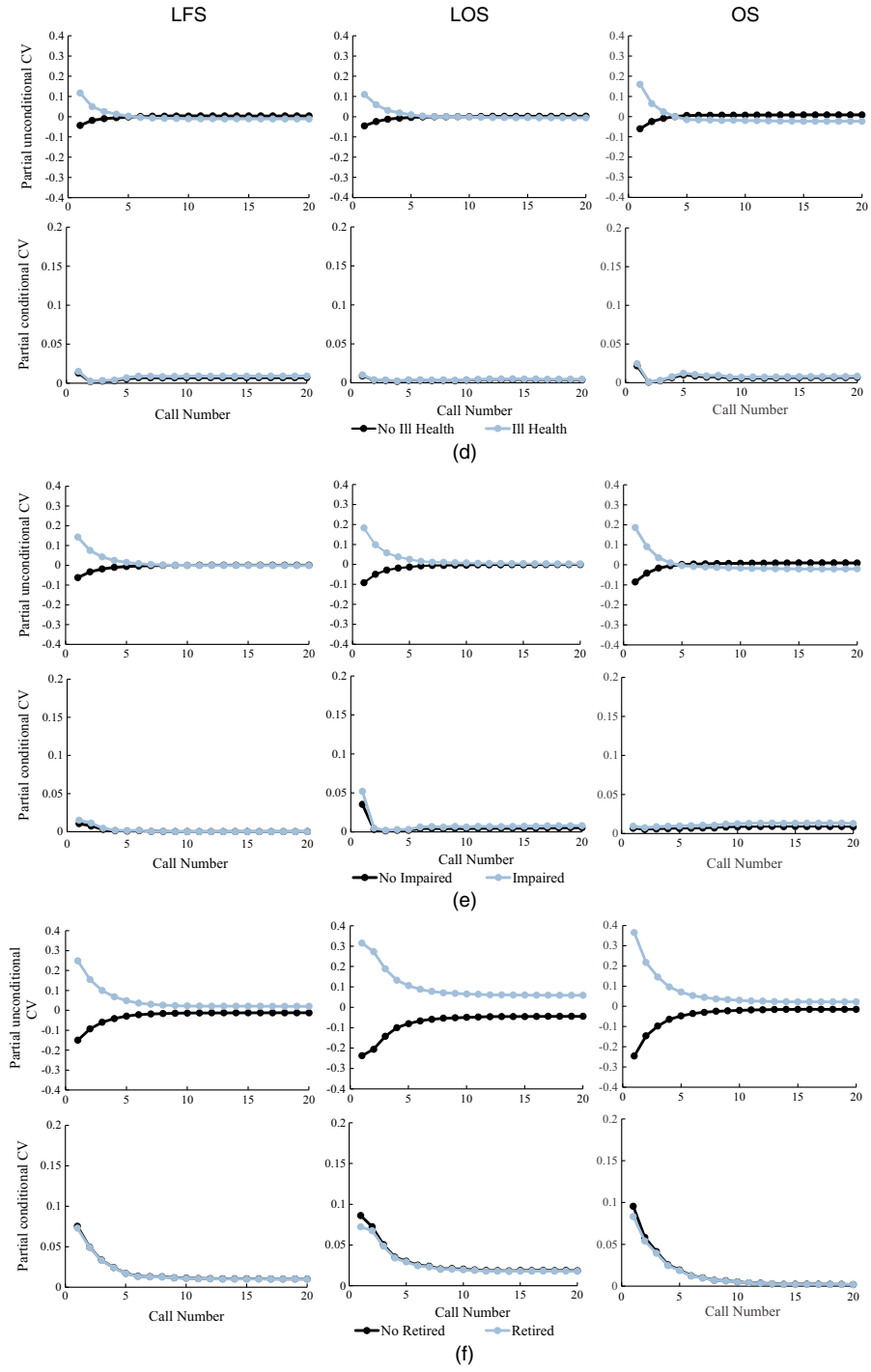

Fig. 3. Unconditional and conditional covariate category CVs over call records in surveys given auxiliary covariate sets including all 10 household attribute covariates, for (a) 'HH economic status', (b) 'HH structure', (c) 'Located in London/SE', (d) 'Ill health individual in HH', (e) 'Impaired individual in HH' and (f) 'Retiree in HH' (each column of graphs details indicators for a survey)

Table 5. Overall and partial unconditional covariate CV PC points in surveys given 'after' and 'during' identification rules and three rule thresholds *a*†

Covariate	*Results for a = 0.01*		*Results for a = 0.02*		*Results for a = 0.05*	
	After	*During*	*After*	*During*	*After*	*During*
LFS						
Overall	8 (2.9%)	7 (4.7%)	6 (7.6%)	5 (12.2%)	4 (19.0%)	4 (19.0%)
HH economic status	8	7	6	5	4	4
HH structure	7	2	1	2	1	2
Retiree in HH	8	7	6	6	5	4
Located in London/SE	5	5	4	4	3	3
LOS						
Overall	11 (9.5%)	8 (15.2%)	8 (15.2%)	7 (18.2%)	6 (22.0%)	5 (27.0%)
HH economic status	9	8	7	6	6	5
HH structure	7	6	6	6	4	3
Retiree in HH	10	8	8	7	6	5
Located in London/SE	6	3	4	3	3	2
OS						
Overall	8 (6.8%)	7 (9.7%)	6 (13.4%)	6 (13.4%)	5 (18.8%)	4 (26.4%)
HH economic status	9	7	7	6	5	5
HH structure	7	4	5	4	3	3
Retiree in HH	10	7	8	7	6	5
Located in London/SE	3	4	3	4	2	2

†The indicator auxiliary covariate sets include all 10 household attribute covariates. For overall CV points, we also present (in parentheses) the percentage of the total calls made to the sample saved by ending collection after the call.

4. Summary and discussion

We address questions concerning the use of representativeness indicators to monitor survey non-response bias risks. We utilize a data set linking paradata detailing attempts to interview sample households in three UK surveys to census household attribute information. The surveys are the LFS, the LOS and the OS. Indicators quantify sample-estimated response propensity variation given an attribute covariate set, with low levels implying representativeness and low non-response bias risks. They are decomposable to measure variation that is associated with covariates and so can inform modifications to data collection methods to improve quality and/or reduce costs. Survey agencies are increasingly interested in utilizing these techniques to manage non-response bias, but guidance on their use is limited, especially concerning monitoring during data collection.

To begin with, indicators require attribute covariates for all sample units: response propensities are statistically modelled. For the first time, we use linked census data: in the UK the only source of information for non-longitudinal surveys. These data are of great utility in our non-response bias analyses. Household linkage rates are around 94%, so the majority of samples can be analysed (though without non-linked household data we cannot completely rule out selection biases). The available covariate set is rich, and samples are from within 3 months of the census, so information will be mostly accurate at the time of survey sampling. Concerning guidance to survey practitioners though, such timeliness is also why we advise caution before using census data more widely for this purpose. The UK census is decadal. How household linkage rates and covariate accuracy decrease for samples further from the census date, reducing data source utility, is unknown. To investigate this, surveys from these dates must be linked.

We also consider indicator use to monitor data collection. First, *R*-indicators can suggest that representativeness is highest early in call records because possible response propensity variation is limited at low response rates. CVs have potentially superior properties as they are less likely to be similarly affected by response rates and also quantify maximum survey estimate mean absolute standardized bias when non-response correlates maximally to the auxiliary covariate set utilized (Schouten *et al*., 2009). We compare indicators in surveys. *R*-indicators behave as described, but CVs suggest that representativeness increases at decreasing rates over call records. This implies that inferences from *R*-indicators are indeed affected by response rates, so we base further explorations on CVs. A barrier to this previously has been that they were less decomposable, but recently partial variants have been presented (de Heij *et al*., 2015; Correa *et al*., 2016). We present approximate partial covariate CV standard errors, by extending the use of the overall CV error approximation of de Heij *et al*. (2014). Unlike similar errors that were derived by de Heij *et al*. (2015), by approximating from the partial *R*-indicator error, our estimators are sometimes not inflated (see Section 2.2 for details). More generally, comparable differences in indicator behaviour arise in other surveys (Lundquist and Särndal, 2013; Correa *et al*., 2016). Hence, concerning guidance to survey practitioners, now that similar functionality exists we recommend that CVs are used to monitor response over call records and in other scenarios where paradata on data collection over time are available (such as mail in–mail back surveys and Web surveys).

Second, there are issues specifying indicator auxiliary covariate sets for use over call records. The same set must be fitted to data at each call for indicators to be informative. Only propensity correlates should be included; otherwise accuracy is affected, but for a given sample size model selection should also lead to reduced covariate retention at low response rates (Schouten *et al*., 2009; Shlomo *et al*., 2012). To advise on selecting covariate sets given such considerations, we study covariate retention across calls and misspecification effects (excluding available correlates, and including non-correlates) on indicators. Regarding covariate retention, in the LFS and LOS fewer are retained in sets given early call data sets than end-of-collection data sets, as predicted. However, in the OS the opposite occurs, and also some LOS covariates are only retained given the early data set. These latter results occur, as covariate (category) partial CVs show (see Section 3.3 and also below), because eventually households in underrepresented categories are interviewed and category response propensities equalize. Such relationships probably often arise in surveys and mean that correct specification of covariate sets (including only correlates) at different calls may vary because of changes in covariate effects as well as the response rate. This makes it even more difficult to choose sets that are not misspecified over parts of the call record.

Regarding set misspecification, exclusion of correlates should lead to comparative overestimation of representativeness, non-correlate inclusion to the opposite and inflated errors. Indicators given the sets above (and sets with all 10 covariates) at calls when sets are identified and correlates known are mostly consistent with these predictions. Differences between sets are small, and CV 95% CIs mainly overlap. Effects are larger given correlate exclusion. Partial CVs suggest that substantial effects on representativeness are underrepresentation of less contactable households (all employed households, no retiree, ill health and impaired individual households, which are overlapping groups), which declines over call records, of 'HHs in London/SE' and of single-adult households. Covariate set differences vary (mostly again being small, though often 95% CIs do not overlap), but excluded correlate unconditional CVs are underestimated, and included non-correlate conditional CVs underestimated. Concerning guidance to survey practitioners, we hence recommend that all available covariates are included in sets that are used to estimate response propensities. Any set is likely to be misspecified over part of the call record, but effects on indicators are mainly small and larger if correlates are excluded (overall representativeness is relatively more overestimated, partial unconditional CVs, which are used to

identify associations then investigated with conditional forms, are underestimated). Therefore, there will be little gain in excluding non-correlates from sets (notwithstanding underestimated conditional covariate effects), and potentially costs since in the process sometime correlates may be excluded.

In addition, we study design PC points, when current methods lead to minimal further increases in quality (or decreases) and modifications should be considered (e.g. Groves and Heeringa (2006)). We identify CV-stability-based points compared with best values over call records ('after' rules), and also previous call values ('during'), with rule thresholds of 0.01–0.05. Partial CV points for covariates linked to substantial effects on representativeness (see earlier for details) differ from overall CV points and also between (non-correlated) covariates. This is to be expected given that they measure different inequalities. In applications, we recommend that overall CV points are used to identify when PC is reached if collection is to be ended completely, as they reflect overall quality. Partial points like those described (and at the category level) are more of interest when modifying methods to improve quality. Effects identified are likely targets as reducing underlying inequalities will lead to the largest improvements (for approaches to using such results to design modifications, see Schouten *et al*. (2012) and Schouten and Shlomo (2016)). In this context, sometimes PC decisions may best be based on these points (e.g. if quality decreases), and/or targeted groups may be treatable separately (see also Groves and Heeringa (2006) and Schouten *et al*. (2013)).

Identified overall PC points range from calls 4 to 11, being earlier in call records as rule thresholds increase. This suggests that in the surveys studied collection (currently up to 20 calls) can indeed be ended early with limited increases in non-response bias risks. Of note to survey agencies that are interested in utilizing these methods to manage risks, call savings made by ending collection at such points compared with sample totals analysed range from 7% to 18% when thresholds equal 0.02 (and increase with threshold size). As well, 'after' points, so named because they are identifiable after collection to inform future periods, tend to be later in call records than 'during' points, which are identifiable during collection as in situations when no historic information exists (e.g. Groves and Heeringa (2006)). This is due to small CV decreases arising from the last responses obtained, which given CV derivation occur even if propensity variation remains similar (see also Lundquist and Särndal (2013)). Practically, such a finding means that 'during' rules identify points at CV values that decrease further with continued effort than 'after' rules: a detail to be considered when the availability of information is an issue.

Finally, we compare PC points across surveys, to provide guidance on whether they can be generalized from one survey to others. This is appealing to survey agencies given issues linking sample information and monitoring costs. We find that LOS overall CV points are one to two calls later than LFS and OS points. Covariate partial CV points are broadly similar. This suggests that generalization could be difficult, even when, as here, surveys are of the same sample frame (some differences between analysed samples exist but do not affect conclusions: see Section 3.4). If LFS or OS points are used, LOS data collection will not achieve the CV stability desired. If LOS points are used, LFS and OS collection will not be optimally efficient. As well, without complete knowledge, errors cannot be identified. Consequently, though we again note the potential benefits of employing these techniques when monitoring data collection in a given survey, we end by recommending that confirmatory work is undertaken before generalizing PC points from one survey to another.

Acknowledgements

This research was funded by the Economic and Social Research Council National Centre for

Research Methods, 'Workpackage1' (grant ES/L008351/1) and the Economic and Social Research Council Administrative Research Centre for England (grant ES/L007517/1). This work contains statistical data from the Office for National Statistics which is Crown copyright. The use of the Office for National Statistics statistical data in this work does not imply the endorsement of the Office for National Statistics in relation to the interpretation or analysis of the statistical data. This work uses research data sets which may not exactly reproduce national statistics aggregates.

References

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*, 2nd edn. Berlin: Springer.

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. *Technical Paper*. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton.

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**, 1029–1049.

Durrant, G. B., D'Arrigo, J. and Steele, F. (2013) Analysing interviewer call record data by using a multilevel discrete time event history modelling approach. *J. R. Statist. Soc.* A, **176**, 251–269.

Durrant, G. B., Groves, G., Staetsky L. and Steele, F. (2010) Effects of interviewer attitudes and behaviours on refusal in household surveys. *Publ. Opin. Q.*, **74**, 1–36.

Durrant, G. B. and 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.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.

Groves, R. M. and Heeringa, S. G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. R. Statist. Soc.* A, **169**, 439–457.

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

de Heij, V., Schouten, B. and Shlomo, N. (2014) RISQ manual 2.0: tools in SAS and R for the computation of R indicators and partial R indicators. (Available from `www.risq-project.eu`.)

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. and 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.

Kappelhof, J. W. S. (2014) The effect of different survey designs on nonresponse in surveys among non-western minorities in The Netherlands. *Surv. Res. Meth.*, **8**, 81–98.

Kreuter, F. (2013) Facing the non-response challenge. *Ann. Am. Acad. Polit. Socl Sci.*, **645**, 23–35.

Luiten, A. and Schouten, B. (2013) Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *J. R. Statist. Soc.* A, **176**, 169–189.

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

Office for National Statistics (2011a) *LFS User Guide*, vol. 1, *Background and Methodology*. Newport: Office for National Statistics. (Available from `http://discover.ukdataservice.ac.uk/catalogue/?sn=6782&type=Data%20catalogue`.)

Office for National Statistics (2011b) Opinions Survey—Technical Report—March 2011. Office for National Statistics, Newport. (Available from `http://discover.ukdataservice.ac.uk/catalogue/?sn=7166&type=Data%20catalogue`.)

Office for National Statistics (2014a) Life Opportunities Survey Waves 1-2, 2009-2012: ONS report. Office for National Statistics, Newport. (Available from `http://discover.ukdataservice.ac.uk/catalogue/?sn=6653&type=Data%20catalogue`.)

Office for National Statistics (2014b) 2011 census variables: part 1. Office for National Statistics, Newport. (Available from `http://www.ons.gov.uk/census`.)

Olson, K. (2006) Survey participation, nonresponse bias, measurement error bias, and total bias. *Publ. Opin. Q.*, **70**, 737–758.

Ouwehand, P. and Schouten, B. (2014) Measuring representativeness of short term business statistics. *J. Off. Statist.*, **30**, 623–649.

Parry-Langdon, N. (2011) Social survey non-response update. *Technical Report*. Office for National Statistics, Newport. (Available from `http://www.ons.gov.uk/ons/dcp171766_240879.pdf`.)

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010) Reduction of nonresponse bias in surveys through case prioritization. *Surv. Res. Meth.*, **4**, 21–29.

Phipps, P. and Toth, D. (2012) Analysing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Ann. Appl. Statist.*, **6**, 722–794.
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.
Roberts, G., Rao, J. N. K. and Kumar, S. (1987) Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1–12.
Särndal, C.-E. and Lundquist, P. (2014) Balancing the response and adjusting estimates for nonresponse bias: complementary activities. *J. Soc. Fr. Statist.*, **155**, 28–50.
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.
Schouten, B., Calinescu, M. and Luiten, A. (2013) Optimizing quality of response through adaptive survey designs. *Surv. Methodol.*, **39**, 29–58.
Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.
Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *J. R. Statist. Soc.* A, **179**, 727–748.
Schouten, B. and Shlomo, N. (2016). Selecting adaptive survey design strata with partial R-indicators. *Int. Statist. Rev.*, to be published.
Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for monitoring and improving representativeness of response. *J. Off. Statist.*, **27**, 231–253.
Shlomo, N., Skinner, C. J. and Schouten, B. (2012) Estimation of an indicator of the representativeness of survey response. *J. Statist. Planng Inf.*, **142**, 201–211.
Steele, F. and Durrant, G. (2011) Alternative approaches to multilevel modelling of survey noncontact and refusal. *Int. Statist. Rev.*, **79**, 70–91.
Wagner, J. R. (2008) Adaptive survey design to reduce nonresponse bias. *PhD Dissertation*. University of Michigan, Ann Arbor.
Wagner, J. R. (2012) A comparison of alternative indicators for the risk of nonresponse bias. *Publ. Opin. Q.*, **76**, 555–575.
Wagner, J. R. and Raghunathan, T. E. (2010) A new stopping rule for surveys. *Statist. Med.*, **29**, 1014–1024.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Online appendix'.