DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 753–779, http://dx.doi.org/10.1515/JOS-2017-0035

Using Response Propensity Models to Improve the Quality of Response Data in Longitudinal Studies

Ian Plewis[1] *and Natalie Shlomo*[1]

We review two approaches for improving the response in longitudinal (birth cohort) studies based on response propensity models: strategies for sample maintenance in longitudinal studies and improving the representativeness of the respondents over time through interventions. Based on estimated response propensities, we examine the effectiveness of different re-issuing strategies using Representativity Indicators (R-indicators). We also combine information from the Receiver Operating Characteristic (ROC) curve with a cost function to determine an optimal cut point for the propensity not to respond in order to target interventions efficiently at cases least likely to respond. We use the first four waves of the UK Millennium Cohort Study to illustrate these methods. Our results suggest that it is worth re-issuing to the field nonresponding cases from previous waves although re-issuing refusals might not be the best use of resources. Adapting the sample to target subgroups for re-issuing from wave to wave will improve the representativeness of response. However, in situations where discrimination between respondents and nonrespondents is not strong, it is doubtful whether specific interventions to reduce nonresponse will be cost effective.

Key words: Representativity indicators; ROC curves; nonresponse; Millennium Cohort Study.

1. Introduction

A longitudinal study needs to retain sample members over time in order to remain representative of its target population, and also to be sufficiently powered so that inferences about measures of change and their correlates are reliable. Intelligent strategies for sample maintenance are required but they need to be based on the efficient use of limited resources. As Couper and Ofstedal (2009, 184) put it in their discussion of keeping in contact with (i.e., tracking) sample members who change address: ". . . time and resources are not limitless, and even if the proportion of nonlocated units can be minimized it may be costly to do so." There is, as yet, little guidance that managers of longitudinal studies can draw on to enable them to optimally allocate the resources available for sample maintenance to reduce unit nonresponse and attrition. This allocation can be done in two ways: within operations in terms of, for example, different ways of minimising the number of refusals, and between operations such as tracking on the one hand and maintaining cooperation on the other.

[1] Social Statistics, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK. Emails: ian.plewis@manchester.ac.uk and natalie.shlomo@manchester.ac.uk.
Acknowledgments: This work was supported by the UK Economic and Social Research Council [RES-175-25-0010 to I.P.].

Emerging findings from research on adaptive survey designs, that is, targeting different sample maintenance strategies at prespecified subgroups of the relevant sample, are beginning to address this gap in the literature. Lynn (2017) reviews this research, particularly as it relates to longitudinal studies. As he points out, the potential of adaptive designs to reduce nonresponse in longitudinal studies is high, especially after the first wave, because it is then possible to draw on data from Wave 1 (and later waves) to generate response propensity models that discriminate between respondents and nonrespondents and which can be used to determine relevant subgroups for targeting. We also use estimated response propensity models to aid decision making and hence our article is related to previous research on adaptive survey designs. Our approach is, however, somewhat different in that we focus on two general research questions:

i) The effectiveness of strategies that are widely, routinely, and nonselectively adopted by survey managers to try to retain sample members at a subsequent wave with the expectation that these strategies will increase precision and reduce attrition bias,
ii) An optimal method of selecting sample members to be the targets of interventions for obtaining a response, for example through refusal conversion.

We refine the first research question by examining the effects on sample representativeness of re-issuing, that is, returning or not returning to the field at waves $t + k$ ($k > 0$) cases who did not respond at wave t ($t \geq 1$). Previous research (Calderwood et al. 2016) using the same data set (the UK Millennium Cohort Study described in Section 3) has examined the effectiveness of re-issuing refusals at wave t who do not initially cooperate at wave t (within wave re-issuing). The effectiveness of re-issuing at the current wave refusals from a previous wave has not, to our knowledge, been studied. The second research question becomes relevant when survey managers consider that routine strategies such as re-issuing refusals at a later wave do not, on their own, sufficiently reduce sample loss and so they deem that further action is needed. Hence, we consider interventions known (ideally from experimental research) to reduce refusals, initially in terms of how well we can predict which cases will refuse and then on the costs of applying the intervention to prevent refusal. This means that we take account of how well the chosen response propensity model discriminates between respondents and nonrespondents rather than the more usual approach of relying solely on the rank order of these response propensities. Our particular approaches to these two general research questions can be extended to decisions about resource allocation for other aspects of sample maintenance and representativeness in longitudinal studies.

The article proceeds as follows. Approaches for preserving the quality of response in longitudinal surveys in terms of sample maintenance strategies and effectiveness of interventions are reviewed in Section 2. Section 3 describes the longitudinal (cohort) study used in our analysis – the UK Millennium Cohort Study. Section 4 sets out our methods and covers response propensity models and two sets of measures derived from them; one set based on the variability of the predicted probabilities of responding (R-indicators) and the other on Receiver Operating Characteristic (ROC) curves. Results of assessing sample maintenance strategies and the effectiveness of possible interventions are presented in Section 5. The article concludes by discussing the implications of the findings and the challenges they present to some of the assumptions made about how best to conduct longitudinal studies.

2. Preserving the Quality of Response in Longitudinal Surveys

In this section we review recent literature on sample maintenance strategies and maintaining response rates. Suppose that our longitudinal studies of interest are based on an initial probability sample drawn from a national population, and that best practice is followed to obtain a good sample at the first wave. Thereafter, cases are lost by dint of being untraced, not contacted, or not cooperating. Managers of such studies can choose to assign their limited resources to (a) locating sample members who change address; (b) reducing noncontact conditional on location; (c) increasing cooperation conditional on contact by:

(1) Maintaining regular contact with sample members over time by, for example, using administrative records. As Couper and Ofstedal (2009) point out, sample members who move are, by definition, different from those who remain at the same address and are also likely to have different patterns of change on variables of interest. Hence, the resources used to locate mobile sample members ought to have both short and longer-term pay offs in terms of reduced bias and increased precision for estimates of change. McGonagle et al. (2009) and Fumagalli et al. (2013) have tested the efficacy of different tracking procedures in the context of long-running household panel surveys.

(2) Minimising noncontact by, for example, careful scheduling of call-backs that draws on field intelligence from previous waves. The issues here are similar to those met in cross-sectional surveys (Durrant et al. 2011).

(3) Maximising cooperation by offering incentives to respondents, by maintaining continuity of interviewers over time, by tailoring between-wave reports to subgroups of the sample, or by providing incentives to interviewers to attain cooperation from cases thought likely by the agency to refuse. In their summary of the evidence about the value of incentives to respondents, Laurie and Lynn (2009) conclude that they do improve response rates but do not necessarily reduce nonresponse bias. Moreover, incentives are expensive and can suffer from the 'deadweight' problem in that they are given to sample members who would have responded without them. Lynn et al. (2014) suggest that interviewer continuity is unlikely to make a substantial difference to response rates and there appears to be little evidence about the extent to which it reduces bias. Fumagalli et al. (2013) tested different ways of reporting back to young and 'busy' respondents in the British Household Panel Survey and found statistically significant but small effects in terms of increased cooperation. Peytchev et al. (2010) randomly allocated cases with an above average propensity not to participate at Waves 4 or 5 of an annual panel survey to interviewers who were either offered or not offered a bonus to secure an interview at the subsequent wave. They found that the monetary incentives offered neither improved response rates nor reduced nonresponse bias for this particular survey. There are, however, other ways in which interviewing resources might be used more efficiently, for example by allocating more experienced interviewers to respondents at greater risk of not cooperating or targeting resources to subgroups that provide the largest contribution to nonresponse bias.

(4) Maintaining a good response rate over time by re-issuing to the field at later waves cases that did not respond, for whatever reason, at earlier waves. As Watson and

Wooden (2014) explain, this practice varies from study to study with implications for response rates and possibly for bias.

(5) Improving response at the current wave by re-issuing (usually to a different interviewer) cases who did not cooperate initially. Calderwood et al. (2016) describe an experiment built into the fourth wave of the UK Millennium Cohort Study where this strategy did improve response rates and also reduced nonresponse bias for some variables. On the other hand, the strategy was expensive as less than a quarter of the re-issued cases were converted into respondents.

We can divide these sample maintenance strategies into two broad categories that correspond to the two general research questions posed in the Introduction: (i) those that are part of the craft and standard practice of managing longitudinal studies such as tracking mobile households and re-issuing refusals and (ii) those that test specific interventions such as offering incentives to interviewers. Strategies within both broad categories should, however, be subject to rigorous assessment of their effectiveness as we illustrate in Section 5.

Raising response rates increases the precision of estimates of the change parameters that are of particular interest to researchers using longitudinal data. If, however, emerging findings from cross-sectional surveys (e.g., Groves 2006) are relevant to longitudinal surveys, then resources dedicated to raising response rates will not necessarily lead to a reduction in bias in estimates of change compared with those obtained from a longitudinal sample based on a lower response rate. Thus, ideally, we would like to be able to balance the costs and benefits of achieving a particular response rate at any wave of a study (and its implications for future waves and thus for measures of change) against the costs and benefits of a lower response rate but where this smaller achieved sample more closely mirrors the longitudinal target population leading to lower nonresponse bias for measures of change. Consequently we need:

i. Evidence about the efficacy of different strategies to reduce the different types of nonresponse, in terms of increasing the response rate and reducing nonresponse bias.
ii. Information about the actual costs of implementing these strategies.
iii. A comparative assessment of the costs of different strategies which, in turn, requires a valuation to be put on the benefits either of a reduction in Mean Square Error (MSE) or separately of increasing precision and reducing bias. This issue has received little attention in the survey literature. Consequently, arguably the best we can do at this stage is to consider the implications for interventions of a range of potential costs and benefits depending on the misclassification errors. In particular, we focus on the ratio of the cost of intervening unnecessarily versus the cost of failing to intervene after allowing for the cost of applying the intervention appropriately.

3. Millennium Cohort Study

We use data from the UK Millennium Cohort Study (MCS), the population for which is babies born in the UK over a twelve-month period during the years 2000 and 2001, and still alive and living in the UK at age nine months. As practically all mothers of new-born

babies in the UK were, at that time, eligible to receive Child Benefit, the Child Benefit register was used as the sampling frame. The Wave 1 sample includes 18,818 babies in 18,552 families living in selected UK electoral wards. The initial response rate at Wave 1 was 72% with response rates by country ranging from 74% in England, 75% in Wales, 76% in Scotland, and 66% in Northern Ireland. Areas with high proportions of Black and Asian families, disadvantaged areas, and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered by electoral ward as described in Plewis (2007b). The standardised design weights vary from 2.0 (England advantaged stratum) to 0.23 (Wales disadvantaged stratum). The first six waves took place when the cohort members were (approximately) nine months, 3, 5, 7, 11, and 14 years old. We focus on Waves 1 to 4 here. Face-to-face interviewing was used, partners were interviewed whenever possible and data were also collected from the cohort members themselves and from their older siblings.

We focus on representativity with respect to Wave 1 for assessing strategies to reduce nonresponse. Plewis (2007b, Sect. 11) shows that survey weights that include adjustments for nonresponse based on auxiliary variables from the Child Benefit register are very similar to the design weights at Wave 1. This, along with the relatively high response rates at Wave 1, leads us to expect good representativity of the target population (the population of births in 2000 and 2001) at the first wave of the study and therefore we use (weighted) Wave 1 as a 'proxy' for the target population.

It has been standard practice in MCS to re-issue to the field all eligible cases at wave t, conditional on their being in the observed sample at Wave 1. Cases become ineligible by virtue of emigration or child death. There are some exceptions to this practice: 'hard' refusals (cases that are fundamentally uncooperative) were never re-issued, and the majority of eligible cases that did not respond at both Waves 2 and 3 were not re-issued at Wave 4.

4. Response Propensity Modelling

We first define response propensity models and then demonstrate their use for addressing our research questions as presented in the Introduction. The predicted probabilities of responding (and their complements) generated by the models can be used in a number of ways. For example, their standard deviations can be used to construct quality indicators, R-indicators, as shown by Schouten et al. (2009) and elaborated in Subsection 4.1. In addition, they can be used to assess the accuracy of discrimination between, or prediction of responding and not responding using ROC curves and logit rank plots as discussed in a survey context by Plewis et al. (2012). The ROC approach is extended in Subsection 4.2 to examine the most efficient way of targeting interventions.

There are many instances in the literature of studies that have modelled the predictors of nonresponse in longitudinal surveys mostly to generate nonresponse weights: for example, Behr et al. (2005); Hawkes and Plewis (2006); Watson and Wooden (2009) and, for MCS, Plewis (2007a), and Plewis et al. (2008). A defining characteristic of these response propensity models is that a binary or categorical outcome of the data collection process – for example, response vs. nonresponse – is linked to a set of explanatory variables, using

either a logit or probit link (or their multivariate equivalents). A simple example is:

$$\text{logit}(\rho_i) = \sum_{k=0}^{K} \beta_k x_{ki} \tag{1}$$

where $\rho_i = E(r_i)$ is the probability of responding for unit i ($i = 1, \ldots, n$); $r_i = 0$ for nonresponse and 1 for response, and x_k are explanatory variables ($x_0 = 1$). ML estimates of β_k ($=b_k$) are easily obtained, leading to predicted probabilities or propensities of responding $\hat{\rho}_i$ where

$$\hat{\rho}_i = e^{\sum b_k x_{ki}} \Big/ \left(1 + e^{\sum b_k x_{ki}}\right) \tag{2}$$

There are different ways of specifying models such as (1) both in terms of which explanatory variables to include and exactly how the response indicator r should be defined. Hence, it will be important to establish whether conclusions are robust to choice of model.

With respect to the UK Millennium Cohort Study (MCS), we consider three models that are used to predict response behaviour in MCS after Wave 1 and which are increasingly complex in terms of their explanatory variables. The specification of the first model is essentially the one used in Plewis (2007a); the explanatory (or auxiliary) variables are all measures obtained at Wave 1 of MCS and are listed in Appendix 2. The second model includes an additional variable that became available from survey managers after Wave 1 and is described in Plewis et al. (2008): whether the main respondent changed address between Waves 1 and 2 and the interactions of this variable with tenure and type of accommodation. The estimates from the first two models allow for the sample design in terms of its disproportionate stratification and clustering using the procedure 'Proc Survey Logistic' in SAS (SAS Institute Inc. 2011). This procedure uses the usual iterative algorithm to compute ML estimates of the regression coefficients for a logistic regression model, but the variance estimation is based on a Taylor expansion approximation under the sample design information. For example, under stratified sample designs, variances are calculated within each stratum separately and then pooled to obtain the overall variance estimates. We show the parameter estimates, standard errors and p-values for Model 2 in Appendix 2 for Wave 2 along with the goodness of fit.

The third model explicitly includes aspects of the sample design: the nine strata and their interaction with the change of address variable as fixed effects, and the primary sampling units which are introduced into the model as a random effect (i.e., a random intercept) so that we have a two level model (main respondents within electoral wards). The third model is:

$$\text{logit}(\rho_{ij}) = \sum_{k=0}^{K} \beta_k x_{kij} + \sum_{p=1}^{P} \gamma_p z_{pj} + u_j \tag{3}$$

where:

ρ_{ij} is the probability of responding for respondent i ($i = 1, \ldots, n_j$) in cluster (i.e., electoral ward) j ($j = 1, \ldots, J$);

x_{kij} are the individual level explanatory variables;
z_{pj} are cluster level dummy variables defining the nine strata;
u_j are normally distributed random effects at level two with mean zero, representing residual variability between clusters.

This model, which is essentially the same as that used in a related context by Durrant and Steele (2009), was estimated using Markov chain Monte Carlo (MCMC) methods available in the *MLwiN* software (Rasbash et al. 2009; Browne 2009), based on 40,000 iterations following a burn-in of 5,000, with noninformative priors throughout. This number of iterations is only half the number used by Durrant and Steele (2009) but, as a result of using orthogonal parameterisation and parameter expansion as described by Browne et al. (2009), convergence was good.

The predicted values $\hat{\rho}_{ij}$ from Equation (3) include the Bayes estimates $\hat{u}_j$ – the means of the 40,000 MCMC iterations – estimating the deviation from expectation of the proportion of response for each cluster. Hence:

$$\hat{\rho}_{ij} = e\left(\sum_{k=1}^{K} b_k x_{kij} + \sum_{p=1}^{P} c_p z_{pj} + \hat{u}_j\right) \Bigg/ \left[1 + e\left(\sum_{k=1}^{K} b_k x_{kij} + \sum_{p=1}^{P} c_p z_{pj} + \hat{u}_j\right)\right] \qquad (4)$$

where b_k and c_p are estimates of β_k and γ_p respectively.

Skinner and D'Arrigo (2011) caution against using the multilevel approach if nonresponse is cluster-specific nonignorable (i.e., where nonresponse depends on unobserved cluster random effects that are correlated with survey variables of interest) as it can lead to bias in weighted estimates. They suggest using conditional logistic regression, conditioning on the number of nonrespondents in each cluster. However, their simulations show that biases are small when clusters are as large as they are here (mean cluster size = 46) and their approach suffers from the disadvantage of excluding from the analysis those clusters with either zero or 100% response. There were 14 out of 398 clusters with 100% response at Wave 1 in the MCS data, accounting for 322 cases.

4.1. Representativity Indicators

We can gauge the effectiveness of widely used strategies for maintaining longitudinal samples over time and the targeting of specific subgroups by using the quality indicators developed by Schouten et al. (2009); Schouten et al. (2011), known as R-indicators and set out in detail below. These measure the extent to which a survey is representative of the population under investigation, that is, they are measures of the degree to which respondents and nonrespondents in a survey differ from each other, both overall and in terms of variables of particular interest.

Up to now, all the published applications of R-indicators have been to cross-sectional surveys as one means of assessing the potential bias-reducing value of additional call-backs and other strategies to ensure a representative sample, and where the response propensities are estimated from a model that uses auxiliary variables from the sample design and from population registers and other administrative sources. Schouten and Shlomo (2017) discuss how to choose strata based on partial R-indicators where more or

less attention is required in the data collection according to a fixed budget for repeated cross-sectional surveys.

Our intention here is to exploit the more detailed information available from the first wave of a longitudinal design firstly to assess the representativeness of later waves in terms of the Wave 1 sample, and then to assess the different sample maintenance strategies and targeting of nonrespondents in terms of their potential to reduce bias in estimates of interest. Our expectation is that this information will help survey managers to make rational decisions about how to allocate limited resources. We return to this argument in the next section.

The Representativity indicators (R-indicators) are based on the variability in the response propensities for a set of sample units s drawn from a population U. The R-indicator is estimated by:

$$\hat{R}_\rho = 1 - 2\hat{S}_\rho \tag{5}$$

where $\hat{S}_\rho^2 = (N-1)^{-1}\sum_s d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$, $d_i = \pi_i^{-1}$ is the design weight, $\hat{\rho}_i$ is defined in (2) or (4), $\hat{\bar{\rho}}_U = \left(\sum_s d_i\hat{\rho}_i\right)/N$ and N may be replaced by $\sum_s d_i$ if it is unknown.

If there is no variation in the response propensities, implying that the propensity to respond for each case is equal to the overall response rate $\hat{\bar{\rho}}_U$, then $\hat{R}_\rho = 1$ and the sample is deemed to be representative of the population from which it was selected, subject to the important caveat that this is conditional on the fitted response propensity model. $\hat{R}_\rho \approx 0$ when $\hat{S}_\rho$ attains its theoretical maximum of 0.5 and there is maximum variability in the response propensities. Shlomo et al. (2012) show that $\hat{R}_\rho$ is biased but the bias is small for large samples. They also show how to derive standard errors for $\hat{R}_\rho$ under simple random sampling. Appendix 1 shows how standard errors can be estimated under complex survey designs.

Schouten et al. (2011) propose unconditional and conditional partial R-indicators as a means of better understanding representativeness for categorical variables of interest and for the actual categories of those variables. Unconditional partial R-indicators ($R_{p(u)}$) for a variable Z having categories $k = 1, 2, \ldots, K$ show how representativeness varies across this variable and thus provides an indication of where the sample is particularly deficient (or satisfactory). Conditional on the response propensity model, the variable level unconditional partial R-indicator is estimated as the between variance:

$$\hat{R}_{\rho(u)} = \hat{S}_B(\hat{\rho}|Z) \tag{6}$$

where $\hat{S}_B^2(\hat{\rho}|Z) = \sum_{k=1}^K \frac{\hat{N}_k}{N}(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U)^2$, $\hat{\bar{\rho}}_k$ is the average of the response propensity in category k: $\hat{\bar{\rho}}_k = \frac{1}{\hat{N}_k}\sum_{s_k} d_i\hat{\rho}_i$, s_k is the set of sample units in category k, and $\hat{N}_k = \sum_{s_k} d_i$.

At the category level $Z = k$, the unconditional partial indicator is estimated as:

$$\hat{R}_{\rho(u),k} = \hat{S}_B(\hat{\rho}|Z=k)\frac{\left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right)}{\left|\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right|} = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right) \tag{7}$$

The unconditional partial indicator $\hat{R}_{\rho(u),k}$ can be positive (denoting over-representation) or negative (denoting under-representation).

Conditional partial R-indicators measure the remaining variance due to variable Z within subgroups formed by all other remaining variables, denoted by X^-. In other words, the conditional partial indicators tell us whether and how a variable Z and its categories contribute to the explanation of response conditional on the other variables in the model.

In the application based on MCS (Section 5), the design weights represent the disproportionate sampling within strata. Therefore, in the calculation of the formula for the R-indicator in (5) and partial R-indicators in (6) and (7) and dropping the notation for the strata, we replace the population size N with the sample size n and $\hat{N}_k$ is replaced by $\hat{n}_k = \sum_{s_k} d_i$. In addition, $\hat{\bar{\rho}}_U$ is replaced by $\hat{\bar{\rho}}_s = \left(\sum_s d_i \hat{\rho}_i\right)/n$ and $\hat{\bar{\rho}}_k = \frac{1}{\hat{n}_k}\sum_{s_k} d_i \hat{\rho}_i$.

The different R-indicators described here can be used as a guide for allocating resources to one kind of sample maintenance activity rather than to another, and for directing resources to obtain response for sample units with particular characteristics. They do not, however, take directly into account the costs of different strategies and interventions, or of the chances of identifying future nonrespondents. This will be addressed in the next section.

4.2. Receiver Operating Characteristic (ROC) Curves

If we want to go beyond routine strategies for sample maintenance and actually intervene to prevent nonresponse, we again want to target our resources in the most efficient way. One way of doing this is by intervening so that everyone with a nonresponse propensity above a threshold is eligible to receive the intervention and nobody with a nonresponse propensity below the threshold receives it. We can determine this threshold by drawing on ROC curves and their associated statistics.

Plewis et al. (2012) show how ROC curves can be used to discriminate between, or to predict whether cases are more likely to be respondents or nonrespondents. In brief, let '+' and '−' denote predictions of nonresponse and response respectively and define c as any threshold (cut point value). Then: '+' is defined by $(1 - \hat{\rho}_i) > c$ where $\hat{\rho}_i$ is the predicted probability of responding from (2) or (4) and '−' is defined by $(1 - \hat{\rho}_i) \leq c$.

The ROC is the plot of $\mathrm{P}(+|\, r = 0)$ against $\mathrm{P}(+|\, r = 1)$ where r is the binary response indicator defined earlier, that is, the True Positive Fraction (TPF) against the False Positive Fraction (FPF) for all thresholds (cut point values) c. The slope of the ROC for any c is just $P(\hat{\rho} = c|r = 0)/P(\hat{\rho} = c|r = 1)$.

The area enclosed by the ROC curve and the horizontal axis, known as the AUC (area under the curve), is of particular interest and this can vary from 1 (when the model for predicting response perfectly discriminates between respondents and nonrespondents) down to 0.5, the area below the diagonal (when there is no discrimination between the two categories). The AUC can be interpreted as the probability of assigning a pair of cases, one respondent and one nonrespondent, to their correct categories, bearing in mind that guessing would correspond to a probability of 0.5. A linear transformation of AUC (= 2*AUC − 1), often referred to as a Gini coefficient, is commonly used as a more natural measure than AUC because it varies from 0 to 1. See Krzanowski and Hand (2009) for a detailed discussion of how to estimate ROC curves and measures derived from them.

One way of determining the optimum threshold (cut point value) is to minimise a cost function such as the one set out by Pepe (2003, 32) for the overall cost of nonresponse per case (TC):

$$TC = C^{+}_{r=0}.TPF.(1 - \hat{\bar{\rho}}_s) + C^{-}_{r=0}.(1 - TPF).(1 - \hat{\bar{\rho}}_s) + C^{+}_{r=1}.FPF.\hat{\bar{\rho}}_s \qquad (8)$$

The first cost term (i.e., $C^{+}_{r=0}$) on the right-hand side of (8) is the actual cost of intervening when an intervention is indicated, that is, the predicted probability of nonresponse is above the chosen threshold (cut point value) c and when the case would indeed have been a nonrespondent ($r = 0$). The second and third cost terms are misclassification costs arising from (i) failing to intervene when the case would have been a nonrespondent and so the cost comes from a valuation of the increase in bias and loss of precision that this entails and (ii) intervening unnecessarily when the case would have responded; $(1 - \hat{\bar{\rho}}_s)$ is the prevalence of nonresponse as defined previously.

An optimum cut point on the response propensity scale can be determined from *TC* by minimising *TC* with respect to the cut point. This implies (Pepe 2003, 72) that the slope of the ROC curve at the optimum cut point is:

$$O^{*}F \tag{9}$$

where O is the odds of being a respondent (and therefore O is usually greater than one) and $F = C^{+}_{r=1}/\left(C^{-}_{r=0} - C^{+}_{r=0}\right)$. Here, F is the ratio of the actual cost of intervening when there would have been a response without the intervention (the false positives) to the cost of failing to intervene for a nonrespondent (the false negatives) minus the assumed to be smaller actual cost of intervening when the prediction to be a nonrespondent is correct (the true positives). Alternative optima appear in the literature. For example, Krzanowski and Hand (2009, 24) focus solely on the costs of misclassification and so F in (9) is then $C^{+}_{r=1}/C^{-}_{r=0}$ with O remaining unchanged. We return to these issues in the context of our example in Section 5 where we also show how determining an optimum cut point can then be used to assess the potential effect of an intervention on R-indicators.

Ideally, any decisions about how to intervene prior to wave t would be based on a response propensity model for that wave but, of course, the required information on the observed response category is not available until after wave t. Consequently, we have to base our decisions on a model for the outcome at wave $t - 1$ and then make the strong but not unreasonable assumption that the accuracy of this model is not substantially diminished at wave t, that the propensities to refuse are strongly associated across the two waves and that the prevalence of refusing is similar across the two waves.

5. Results from the MCS

We present in this section results from the MCS based on the proposed methods in Section 4 using R-indicators and ROC curves for assessing representativity of re-issuing strategies and interventions for targeting nonrespondents.

5.1. Representativeness in Re-Issuing Strategies

R-indicators are used here to provide evidence about the utility of different re-issuing strategies after Wave 1. In particular, we compare the strategies that were used by MCS survey managers with a set of alternative re-issuing strategies that might plausibly be adopted in the future by those managers or by managers of other longitudinal studies. The strategies are labelled (i) S (for standard practice) so S2, S3, and S4 refer to the standard practices applied at Waves 2, 3, and 4 in MCS; (ii) P, the hypothetical strategy of only re-issuing to the field productive cases (i.e., cases who provided some but not necessarily

complete data) from previous waves so P3.2 refers to a strategy of only re-issuing at Wave 3 cases that were productive at Wave 2, P4.23 refers to only re-issuing at Wave 4 cases that were productive at Waves 2 and 3, P4.3 to only re-issuing at Wave 4 cases that were productive at Wave 3 (including some that were not productive at Wave 2); (iii) C (for cooperation), the hypothetical strategy of not re-issuing refusals from previous waves so C3.2 refers to a strategy of not re-issuing at Wave 3 refusals from Wave 2, C4.23 refers to not re-issuing at Wave 4 refusals from Waves 2 and 3, C4.3 to not re-issuing refusals just from Wave 3.

As the estimates of representativeness based on the R-indicators were similar for all strategies under the three response propensity models presented in Section 4, we show results just for Model 2 here. Results for other models are available on request. Note that we do use all three models when discussing targeting interventions, as explained in Subsection 5.2. For this model, the calculation of the response propensities, the R-indicator and their standard errors all take into account the complex survey design of the MCS with respect to clustering, survey weights and stratification (see Appendix 1). We also test the significance of the difference between R-indicators of strategies within relevant waves as follows: P3.2 and C3.2 compared with standard practice S3 in Wave 3; P4.23, P4.3, C4.23, and C4.3 compared with standard practice S4 in Wave 4. In order to test for significant differences in the R-indicators of strategies that differ from the standard practice, we estimate the standard error of the difference between R-indicators using 1,000 bootstrap samples with replacement. In each bootstrap sample, we re-estimate the response propensity models for the standard practice and the strategy of interest without taking into account the complex survey design. We calculate the R-indicator for the standard practice and the R-indicator for the strategy of interest and obtain the bootstrap variance of the difference between the R-indicators. This variance is then adjusted to account for the complex survey design based on the original sample.

Table 1 presents the R-indicators and their 95% confidence intervals for standard practices S2, S3, and S4. In addition, we provide the R-indicators and their confidence intervals for the re-issuing strategies as well as the difference between these strategies from the standard practice using the bootstrap standard errors. The number of cases that would have been lost and their percentage of the actual productive sample at each wave as a result of the different re-issuing strategies, all other things being equal, are also shown in this table.

From Table 1, we see that, with respect to Wave 1, the representativity for the standard re-issuing practice in Waves 2 and 3 was very similar but there was a marked decline at Wave 4. For those rows labelled P, we see that representativeness falls if only productive cases from previous waves are re-issued and the R-indicators are all significantly lower than for the standard practice in each wave. The R-indicator for P4.3 (re-issuing to Wave 4 productive cases at Wave 3) is significantly higher than the R-indicator for P4.23 ($p < 0.001$) (re-issuing to Wave 4 productive cases at both Waves 2 and 3) and is closer to the R-indicator of the standard practice S4.

Comparing C3.2 to S3 suggests that representativeness is less compromised if only refusals (rather than all unproductive cases) from previous waves are not re-issued since the R-indicator is more similar to standard practice although still significantly lower. Comparing the strategies of cases that refused just at Wave 3 not being re-issued at Wave 4 (C4.3) to

Table 1. R-indicators (with 95% confidence intervals (CI)) for standard practice and different re-issuing strategies and number of cases lost with the percentage of actual productive sample at each wave.

Strategies	R-indicator (95% CI)	Difference from standard practice (95% CI)	Cases lost	Percentage of actual productive sample
S2	0.781 (0.763–0.799)	-	n.a.	-
S3	0.794 (0.777–0.811)	-	n.a.	-
P3.2	0.715 (0.694–0.735)	−0.079* (−0.094– −0.065)	1,444	9.5%
C3.2	0.771 (0.752–0.789)	−0.023* (−0.031– −0.015)	473	3.1%
S4	0.740 (0.720–0.760)	-	n.a.	-
P4.23	0.666 (0.644–0.687)	−0.074* (−0.088– −0.060)	1,642	12.3%
P4.3	0.715 (0.694–0.736)	−0.024* (−0.032– −0.016)	613	4.6%
C4.23	0.721 (0.700–0.741)	−0.019* (−0.028– −0.010)	525	3.9%
C4.3	0.735 (0.715–0.756)	−0.004 (−0.009– 0.001)	205	1.5%

Notes

1. Data source: MCS, Waves 1 to 4.
2. * – significant difference ($p < 0.05$) from standard practice.
3. Sample sizes after omitting ineligibles (see Section 3): 18,143 (W2), 17,990 (W3), and 17,819 (W4). The size of the productive sample at Wave 1 is 18,552 but a few cases were omitted from the response propensity models because of item nonresponse at Wave 1.

Table 2. Unconditional partial R-indicators (with 95% confidence intervals (CI)) for main respondent's highest educational qualifications.

Strategies	Partial R-indicator (95% CI)	Difference from Standard Practice (95% CI)
S2	0.060 (0.052–0.067)	-
S3	0.060 (0.053–0.067)	-
P3.2	0.083 (0.075–0.092)	0.024* (0.020–0.027)
C3.2	0.067 (0.059–0.074)	0.007* (0.006–0.008)
S4	0.077 (0.069–0.085)	-
P4.23	0.100 (0.091–0.109)	0.023* (0.020–0.026)
P4.3	0.086 (0.078–0.094)	0.009* (0.008–0.010)
C4.23	0.083 (0.075–0.092)	0.007* (0.006–0.008)
C4.3	0.078 (0.070–0.086)	0.002* (0.001–0.002)

Notes

1. Data source: MCS, Waves 1 to 4.
2. * – significant difference ($p < 0.05$) from standard practice.
3. Sample sizes after omitting ineligibles (see Section 3): 18,143 (W2), 17,990 (W3), and 17,819 (W4). The size of the productive sample at Wave 1 is 18,552 but a few cases were omitted from the response propensity models because of item nonresponse at Wave 1.

those that refused at Waves 2 and 3 (C4.23), we see a slight and significant increase in representativeness for C4.3 ($p < 0.001$) and this strategy has similar representativeness compared to the standard practice in S4 with a nonsignificant difference between the R-indicators. In addition, strategy C4.3 has the lowest number of dropped cases.

Tables 2 and 3 give estimates of variable-level unconditional partial R-indicators for two categorical variables that are associated with many of the variables of interest in studies that use MCS: the main respondent's highest educational qualification and ethnic group. Here the higher the partial R-indicator, the more it is contributing to the lack of representative response.

Table 3. Unconditional partial R-indicators (with 95% confidence intervals (CI)) for main respondent's ethnic group.

Strategies	Partial R-indicator (95% CI)	Difference from Standard Practice (95% CI)
S2	0.034 (0.025–0.044)	-
S3	0.031 (0.023–0.039)	-
P3.2	0.047 (0.036–0.058)	0.016* (0.012–0.021)
C3.2	0.033 (0.024–0.041)	0.002 (0.000–0.003)
S4	0.032 (0.023–0.041)	-
P4.23	0.052 (0.039–0.064)	0.020* (0.015–0.024)
P4.3	0.040 (0.030–0.051)	0.008* (0.006–0.011)
C4.23	0.035 (0.025–0.044)	0.003* (0.001–0.005)
C4.3	0.034 (0.024–0.043)	0.002* (0.001–0.003)

Notes

1. Data source: MCS, Waves 1 to 4.
2. * – significant difference ($p < 0.05$) from standard practice.
3. Sample sizes after omitting ineligibles (see Section 3): 18,143 (W2), 17,990 (W3), and 17,819 (W4). The size of the productive sample at Wave 1 is 18,552 but a few cases were omitted from the response propensity models because of item nonresponse at Wave 1.

From Tables 2 and 3, the unconditional variable level partial R-indicators for the two variables across the re-issuing strategies are all significantly different from zero. Thus, the distributions of highest educational qualifications and, to a lesser extent, ethnic group are less representative at Waves 2, 3, and 4 under the standard practice (S2, S3, and S4) when compared with Wave 1. Conditional variable level partial R-indicators (not presented here) are all significantly different from zero showing that the lack of representativity remains even after conditioning on other variables. Similar to Table 1, Tables 2 and 3 also present the confidence interval of the difference in the variable level partial R-indicators based on the bootstrap samples comparing a re-issuing strategy to its standard practice in the respective wave. For the unconditional variable level partial R-indicators, all re-issuing strategies are significantly lower compared to the standard practice in each wave and that only issuing productive cases from Waves 2 and 3 at Wave 4 (P4.23 in Tables 2 and 3) leads to the poorest representation for both variables.

Other variables with high variable level partial R-indicators are banded family income, family status and housing tenure. Out of these variables, the estimates for individual categories (not presented here) indicate over-representation of those with higher rather than lower or no qualifications, are of white British compared with minority ethnic groups, owned their accommodation, family status of two parents and higher income levels.

We can use the information from the categorical partial R-indicator estimates (not presented here) to reassess the re-issuing strategy and targeting of the nonrespondents. For example, considering the re-issuing strategy at Wave 3, we might decide not to re-issue the cases that were unproductive at Wave 2 (row P3.2) only if they belonged to the majority ethnic group or had some educational qualifications. The estimate of the R-indicator then increases from 0.715 (0.694–0.735) to 0.743 (0.724–0.761) for Model 2 with 510 cases lost and so the strategy is closer in effectiveness to the strategy of not re-issuing at Wave 3 refusals from Wave 2 (i.e., C3.2).

Similarly, looking at the re-issuing strategy at Wave 4, we might decide not to re-issue the cases that were unproductive at earlier waves (row P4.23) only if they belonged to the majority ethnic group or had some educational qualifications. The estimate of the R-indicator then increases from 0.666 (0.644–0.687) to 0.689 (0.669–0.709) with 481 cases lost and so the strategy is closer in effectiveness to re-issuing all productive cases from Wave 3 (i.e., P4.3).

5.2. *Use of ROC in Determining Interventions*

Table 4 shows how discrimination between respondents and all nonrespondents, as measured by the Gini coefficient, varies by model and wave. We present estimates for all

Table 4. Gini estimates (with 95% confidence intervals (CI)) by model and wave.

	Model 1	Model 2	Model 3
Wave 2	0.39 (0.37–0.41)	0.39 (0.37–0.41)	0.42 (0.40–0.44)
Wave 3	0.37 (0.35–0.39)	0.37 (0.35–0.39)	0.36 (0.34–0.38)
Wave 4	0.37 (0.35–0.39)	0.37 (0.35–0.39)	0.39 (0.37–0.41)

Notes
1. Data source: MCS Waves 1 to 4.
2. Sample sizes for all models: Wave 2 = 18,148, Wave 3 = 17,990, and Wave 4 = 17,819.

three models in this section as the differences between them are more marked than they were for the R-indicators of the previous section. In all cases, the AUC were estimated assuming a bi-normal model (Pepe 2003), and the Gini coefficients and their standard errors were derived from these estimates. We see that discrimination at Waves 2 and 4 is improved when considering the sample design as a random effect (Model 3) and is generally only slightly reduced at Waves 3 and 4 compared with Wave 2, even though the models are based on Wave 1 variables and so do not allow for changes in these variables between waves.

We now consider how we might target interventions designed to prevent refusals at wave t, conditional on being in the Wave 1 sample. Our interest is in directing interventions known from previous research to have an effect on converting refusals into productive cases (incentives for example), that is, targeting cases most likely to benefit from an intervention in terms of their estimated propensities not to respond. Here we focus on Wave 3 when some data was obtained from 80% of the eligible Wave 1 sample and refusals were nearly twice as common as other kinds of unproductive cases. We also consider how robust the targeting is to misclassification of the outcome variable and to changes in the response propensity model. We consider misclassification by defining refusal in two ways: (i) as recorded by the interviewer and (ii) by including those cases ($n = 166$) that were noncontacts at Wave 2 and refusals at Wave 3 as noncontact can sometimes be a hidden refusal (Blom 2014).

We are also able to improve the discrimination from Model 2 by including an assessment of the neighbourhood by interviewers at the times they called at sample households at Wave 2. This variable, which can be thought of as paradata in Kreuter et al.'s (2010) terms, was not available for the 'not located' group (and so could not be used in Subsection 5.1). It is described in more detail in Appendix 3. Now our dependent variable is 0 for refusal and 1 for productive, with all other nonrespondents omitted from the models (leading to a reduction in sample sizes compared with Table 4). Table 5 presents the relevant Gini estimates and shows that discrimination between refusals and productives now improves across the three models but is somewhat lower for Model 3 for the more encompassing definition of refusal.

We now turn to estimating the optimum thresholds (cut points) that take account of prevalence and costs as set out in Subsection 4.2. The estimates of the odds ratio O from Equation (9) for the type of refusal defined in (i) above are 8.6 (Model 1) and 10.5 (Models 2 and 3). We consider a range of plausible values of the cost ratio F: 0.33, 0.8, and 1.5. The values of $F < 1$ imply that the cost of intervening unnecessarily is less than the cost of failing to intervene after allowing for the cost of applying the intervention appropriately.

Table 5. Gini estimates (with 95% confidence intervals (CI)) by model, Wave 2.

Type of refusal (Prevalence, Wave 2)	Model 1	Model 2	Model 3
Refusal (i) (10.0%)	0.35 (0.32–0.38)	0.40 (0.37–0.43)	0.52 (0.49–0.55)
Refusal (ii) (11.0%)	0.35 (0.32–0.38)	0.41 (0.38–0.44)	0.43 (0.40–0.46)

Notes

1. Data source: MCS Waves 1 and 2.
2. Sample sizes: (a) refusal, Model 1–16,468 and 16,627; (b) refusal, Models 2 and 3 – 15,647 and 15,813.

This is due to the expectation that it will be more costly to lose cases which may have an impact on nonresponse bias. It is not, in fact, possible to estimate a cut point threshold with any confidence when $F = 1.5$, nor when $F = 0.8$ for Models 1 and 2 because they are in the extreme tails of the two distributions. For $F = 0.33$, the optimum cut point values of $(1 - \hat{\rho})$ (the estimated probability of not responding) for the three models are 0.28, 0.32, and 0.23; for $F = 0.8$ for Model 3, the cut point value for $(1 - \hat{\rho})$ is 0.46. Thus, the chosen intervention would be targeted at those cases with a propensity not to respond at least as large as these estimates. Note that optimal values were obtained after applying Box-Cox transformations (Box and Cox 1964) to the conditional distributions required to estimate the slope of the ROC.

Table 6 shows how the optimum sizes of the intervention groups vary by model and, for Model 3, by cost ratio. We see that these sizes are sensitive to model specification and choice of F, ranging from 802 ($F = 0.33$, Model 3) down to 82 ($F = 0.8$, Model 3). Table 6 also shows what proportions of the intervention groups actually refused at Wave 2 (remember that the model for decisions at Wave 3 is, perforce, based on Wave 2 data) and the response outcome for these groups at Wave 3. The final row shows how much larger the sample at Wave 3 would have been if the intervention to convert refusals had a 100% success rate. In practice, of course, the success rate will be much lower.

Similar results to those shown in Table 6 are obtained when the more inclusive definition of refusal (i.e., (ii) above) is used.

Focussing on Model 2 when $F = 0.33$, we can see what the effect on the R-indicator would have been if all the refusals ($n = 73$) had been converted into respondents. We find an increase from 0.771 (0.752–0.789) when no refusals are re-issued (see Table 1 row C3.2) to 0.809 (0.793–0.826) so that representativity would significantly improve if the intervention had a 100% success rate. The results in Table 6 for Model 3 when $F = 0.33$ suggest that this improvement might be even more substantial.

5.3. *Impact on Nonresponse Bias of Banded Income*

A key question is whether we can reduce nonresponse bias by adapting the re-issuing strategy to ensure a more representative set of respondents. To assess the impact on nonresponse bias, we examine a key variable – banded income measured as a six-category variable with category 1 being the lowest income and category 6 being the highest income (but treated as continuous here). We use the banded income as measured at Wave 1 for this comparison. Table 7 presents the unweighted mean income variable at Wave 1 (ignoring

Table 6. Intervention group sizes and outcomes.

	Model 1, $F = 0.33$	Model 2, $F = 0.33$	Model 3, $F = 0.33$	Model 3, $F = 0.8$
Group size (n)	505	249	802	82
Refusal, Wave 2 (%)	29	33	33	52
Refusal, Wave 3 (%)	31	29	25	34
Maximum increase in Wave 3 sample (n)	155	73	201	28

Data source: MCS Waves 1 to 3.

Table 7. Mean income category (unweighted) for Wave 1 and Wave 3 according to re-issuing strategies with standard error (S.E.) and relative absolute bias (RAB).

		Wave 3 Strategies (based on Wave 1 income)				
Ethnic Group	Wave 1	S3	P3.2	P3.2 (targeted non-whites and low education)	C3.2	C3.2 ($F = 0.33$ and 100% refusal conversion)
Total	3.33	3.42	3.47	3.42	3.43	3.41
S.E.	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
RAB	-	2.7%	4.2%	2.7%	3.0%	2.4%
White British	3.41	3.50	3.54	3.54	3.51	3.49
S.E.	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
RAB	-	2.6%	3.8%	3.8%	2.9%	2.3%
Other	2.89	2.95	3.00	2.86	2.95	2.94
S.E.	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
RAB	-	2.1%	3.8%	1.0%	2.1%	1.7%

Data source: MCS Waves 1 to 3.

missing values) under the following scenarios for Wave 3: standard re-issuing practice (S3); re-issuing strategy of only those productive at Wave 2 (P3.2); re-issuing strategy of only those productive at Wave 2 (P3.2) and in addition the targeted non-whites and low educational qualifications; re-issuing strategy of not including refusals from Wave 2 (C3.2); re-issuing strategy based on the intervention shown in Table 6 for Model 2 with $F = 0.33$ and assuming a 100% success rate of refusal conversion.

From Table 7, we see a difference in the mean income as measured in Wave 1 when comparing the standard re-issuing strategy of Wave 3 (S3). Those cases in lower income categories are more likely to drop out of the survey by Wave 3. Comparing the other re-issuing strategies at Wave 3 with Wave 1, the P3.2 strategy shows the largest nonresponse bias and C3.2 following interventions shows the least nonresponse bias for the Total and 'White British'. However, we see a significant reduction in nonresponse bias for the 'Other' ethnic groups under the targeting strategy of P3.2 where the non-whites were specifically targeted. Further refinement of strata for targeting response can lead to higher gains when the target variables of interest are correlated with the explanatory variables of the response propensity models.

6. Conclusions

We return to the two general questions posed in the Introduction, now with the benefit of some evidence. We have seen how survey managers might assess the effectiveness of their sample maintenance strategies by drawing on R-indicators. These provide both an overall assessment, and assessments for particular variables, of the extent to which longitudinal samples maintain representativeness over time. Individual researchers might reasonably argue that they are more interested in information about how bias and precision for particular estimates of change vary according to the strategies adopted, an approach taken by Pudney and Watson (2013). But managers have to balance the competing demands of different research teams and, for them, an overall indicator of representativeness is potentially valuable.

The results in the previous section are, of course, only illustrative, restricted as they are to just one birth cohort study conducted in a particular way, and where waves are more irregularly spaced and less frequent than they usually are in household panel surveys, for example. They need to be replicated on other surveys. Nevertheless, they do provide some pointers about the utility of different sample maintenance strategies and the targeting of sample units. The evidence does support the strategy of re-issuing to the field nonresponding (i.e., unproductive) cases at later waves. However, the gains in terms of representativeness from re-issuing refusals are smaller than for re-issuing all unproductive cases with the implication that more is to be gained from re-issuing cases that were not located or not contacted (with the majority falling into the not located group) than re-issuing refusals, bearing in mind that many of the re-issued refusals continue to refuse at subsequent waves. There is some evidence that targeting sample units for response according to partial R-indicators may reduce nonresponse bias when explanatory variables are associated with the target variables. These conclusions appear to be robust to changes in the specification of the response propensity model. One caveat should be attached to the findings that are based on the R-indicators: it does not

necessarily follow that strategies that maintain representativeness at later waves of a longitudinal study compared with Wave 1 also maintain representativeness with respect to the target population.

The ROC method can help managers to decide whether a particular (i.e., non-routine) intervention to reduce nonresponse is likely to be cost-effective. The main point to emerge from the ROC analyses of the MCS data is to question the value of implementing specific intervention strategies of the kind discussed in Section 2 to reduce nonresponse in longitudinal studies. There are a number of reasons for being doubtful. The predictive accuracy of response propensity models based on variables measured at Wave 1 is not high and hence they do not clearly discriminate between productive and different kinds of unproductive cases leading to higher misclassification costs. This means that the ROC curves are relatively shallow, although less shallow than they would be if based just on auxiliary data from the sample frame. Given that the prevalence of types of nonresponse at any single wave is usually relatively low (and so O in Equation (9) is high), we require the cost ratio F to be small in order to be able to estimate the slope of the ROC curve with some precision. To suppose that the net cost of failing to intervene to prevent a case of nonresponse is three times the cost of intervening unnecessarily (i.e., $F = 0.33$) is arguably too extreme and the value of 0.8 is perhaps more realistic. Clearly, however, this will depend on the nature of the intervention with those involving incentives carrying a greater deadweight. It might only be cost-effective to have an intervention to prevent refusal directed at a very small group. Moreover, this conclusion is predicated on two further assumptions: that any decision to intervene at wave t can be reliably based on a response propensity model generated for the response outcome at wave $t - 1$, and that we have available to us interventions that have been established to be highly effective at reducing different kinds of nonresponse. The results in Table 1 suggest that the first of these assumptions is not unreasonable but the second assumption is a very strong one and not generally supported by the literature.

The conclusions about intervening are dependent on the chosen cost function. This assumes that costs for any two subjects are independent. This might not hold for clustered designs when, for example, the actual cost of intervening for cases in the same cluster might be less than for cases in a different cluster. Also, it would be worth investigating alternatives to the 'all or nothing' policy considered here such as taking a majority of cases with a propensity not to respond above the threshold and a minority from cases with a lower propensity not to respond. This could be justified by the fact that the propensities are estimated and are subject to sampling and specification error. Other decision rules, not based on ROC curves, are also possible. For example, Schouten and Shlomo (2017) target sample units by grouping them in strata based on partial R-indicators and selecting strata for follow-up according to a given threshold depending on costs. Alberman and Goldstein (1970) suggest maximising a utility function that is based on the number of poor outcomes that are prevented subject to the constraint that the available resources are fixed.

This article has indicated how to use response propensity models in a way that can help survey managers to determine the optimal allocation of resources for sample maintenance within longitudinal studies. There are, however, a number of outstanding questions such as how to decide between competing maintenance strategies as well as broader questions

such as the utility of sacrificing some cases in order to, for example, collect more information from each case. These kinds of decisions require managers not only to have accurate cost data but also indications from users of the benefits of different allocations. We do not pretend to have shown how to allocate scarce resources to the different aspects of longitudinal studies. Clearly more research is needed based on a range of longitudinal designs and data collection methods. We do, however, believe that the methods set out here offer a way forward.

Appendix 1

Calculating the Variance of the R-indicator Under Complex Survey Designs

Shlomo et al. (2012) show that the variance of the R-indicator comes from considering the sum of two components:

$$\text{var}\left(\hat{S}_\rho^2\right) = E_{\hat{\boldsymbol{\beta}}}\left[\text{var}_s\left(\hat{S}_\rho^2\right)\right] + \text{var}_{\hat{\boldsymbol{\beta}}}\left[E_s\left(\hat{S}_\rho^2\right)\right] \tag{A1}$$

where the subscript $\hat{\boldsymbol{\beta}}$ denotes the distribution induced by $\hat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ takes into account the complex survey design using the sandwich estimator of Binder (1983) as calculated in the SAS Proc Survey Logistic procedure (SAS Institute Inc. 2011).

For the first term in (A1), following Shlomo et al. (2012) and given the consistency of $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ under standard sampling designs for the expectation, we can estimate the first term by a design-based estimator of $\text{var}_s\left[\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}} u_{hji}\right]$ taking into account the complex survey design where u_{hji} is replaced by $d_{hji}\left(\hat{\rho}_{hji} - \bar{\rho}_U\right)^2$, $\hat{\rho}_{hji}$ is calculated under the SAS Proc Survey Logistic procedure using $\hat{\boldsymbol{\beta}}$, d_{hij} is the design weight of unit i in cluster j and stratum h, m_{hj} is the number of units in cluster j and stratum h and n_h is the number of clusters in stratum h.

The variance is estimated as:

$$\text{var}_s\left[\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}} u_{hji}\right] = \sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(u_{hj.} - \bar{u}_{h..})^2 \tag{A2}$$

where f_h is the sample fraction in stratum h, $u_{hj.} = \sum_{i=1}^{m_{hj}} d_{hji}u_{hji}$ and $\bar{u}_{h..} = \left(\sum_{j=1}^{n_h} u_{hj.}\right)/n_h$.

The second term of (A1) is calculated as described in Shlomo et al. (2012) but we replace the covariance term shown for the case of simple random sampling with the sandwich estimator.

To obtain the estimated variance of the unconditional partial R-indicator at the variable level Z as shown in Tables 2 and 3: $\hat{S}_B(\hat{\rho}|Z) = \sqrt{\sum_{k=1}^{K}\frac{\hat{N}_k}{\hat{N}}(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U)^2}$, we note that $\hat{S}_B^2(\hat{\rho}|Z)$ is the variance of the estimated response propensities when the stratification is on variable Z only. Therefore, we use the calculation of the variance of the R-indicator under complex survey designs as described above where the response propensities are modelled under a logistic regression having a single auxiliary variable Z.

Recall that in the application in Section 4 based on the Millennium Cohort Study, the design weights are calculated relative to the sample in Wave 1 and therefore, the

calculation of all variance estimates are relative to the sample size n and not to the population level N.

Appendix 2

Predictors of Nonresponse and Their Categories for Models Presented in Section 4 (see Table 2 In Plewis (2007a)) with Percentages Based on Wave 2 of the MCS

1. Family income (Six ordered categories: 1.6%; 24.0%; 30.5%; 18.0%; 13.0%; 4.6%, and Not applicable 5.8%; Missing 2.5%)
2. Ethnic group of cohort child (White British (82.8%); Mixed (3.0%); Indian (2.5%); Pakistani/Bangladeshi (6.8%); Black/Black British (3.6%); Other (1.4%))
3. Accommodation type (House (85.0%); other (15.0%))
4. Tenure (Own (58.0%); rent (35.7%); other (6.3%))
5. Main respondent's age (<30 (50.2%); 30 + (49.9%))
6. Main respondent's educational qualifications (None (2.8%); NVQ1-5 (3.3%; 12.3%; 8.4%; 9.3%; 44.4%); other/overseas (19.5%))
7. Cohort child breast fed (Yes: 66.8%)
8. Longstanding illness, main respondent (Yes: 21.0%)
9. Adults in household (1 (17.2%) 2 (71.4%) 3 (1.2%) 4 + (10.2%))
10. Main respondent voted in last general election (Yes: 50.7%)
11. Gave consent to record linkage (Yes: 93.1%)
12. Provided a stable address at Wave 1 (Yes: 82.5%)
13. Whether moved between Waves 1 and 2 (Yes: 38.2%)

We next present in Table A2.1 the model estimates and their standard errors for the logistic regression response model for Wave 2 of the MCS based on predictors as defined for Model 2 according to Wave 1 values and accounting for the complex survey design.

Table A2.1. Estimates for the logistic regression response Model 2 for Wave 2 using predictors from Wave 1 accounting for the complex survey design.

Parameter		Estimate	S.E.	t Value	Pr > |t|	Exp(Est)
Intercept		0.67	0.105	6.44	<.000	1.96
Cohort child breast fed	Yes	0.17	0.028	6.22	<.000	1.19
Long term illness respondent	Yes	0.13	0.033	3.97	<.000	1.14
Adults in household	1	0.12	0.072	1.72	0.086	1.13
	2	0.34	0.064	5.35	<.000	1.41
	3	−0.05	0.174	−0.28	0.783	0.95
Age of mother	<30	−0.15	0.031	−4.87	<.000	0.86
Mother's education	None	0.20	0.124	1.62	0.106	1.22
	NVQ1	0.26	0.085	3.08	0.002	1.30
	NVQ2	0.19	0.086	2.23	0.027	1.21
	NVQ3	0.06	0.081	0.78	0.434	1.07
	NVQ4	−0.07	0.051	−1.44	0.152	0.93
	NVQ5	−0.34	0.114	−2.95	0.003	0.72
Ethnicity	White British	0.23	0.068	3.38	0.001	1.26
	Mixed	0.003	0.117	0.02	0.981	1.00
	Indian	0.23	0.148	1.55	0.123	1.26
	Pakistani/Bangladeshi	0.15	0.109	1.39	0.166	1.16
	Black	−0.20	0.107	−1.86	0.063	0.82
Accommodation	Owns House	0.20	0.035	5.61	<.000	1.22

Table A2.1. Continued.

Parameter		Estimate	S.E.	t Value	Pr > |t|	Exp(Est)
Income	Low	−0.11	0.159	−0.71	0.479	0.89
	2	−0.15	0.064	−2.37	0.018	0.86
	3	−0.01	0.055	−0.26	0.795	0.99
	4	0.19	0.067	2.80	0.005	1.21
	5	0.29	0.091	3.14	0.002	1.33
	High	0.09	0.129	0.70	0.482	1.10
	Not applicable	0.01	0.095	0.16	0.876	1.02
Stable Address	Yes	−0.15	0.029	−5.11	<.000	0.86
Tenure	Own	0.09	0.058	1.58	0.114	1.10
	Rent	−0.05	0.046	−1.15	0.250	0.95
Consent to Linkage	Yes	0.37	0.048	7.78	<.000	1.45
Voted at last election	Yes	0.16	0.028	5.62	<.000	1.17
Mobile: Whether moved between Wave 1 and 2	Yes	−0.08	0.039	−2.09	0.037	0.92
Accommodation*mobile		0.12	0.031	3.94	<.000	1.13
Tenure (Own)*mobile		0.11	0.041	2.80	0.005	1.12
Tenure (Rent)*mobile		0.17	0.042	3.98	<.000	1.18

Wald Statistic $F(34,389) = 41.4$, $p < 0.001$, AIC = 14636.

Note: Data source: MCS, Waves 1 and 2.

Appendix 3

Interviewer Assessments of the Neighbourhood, MCS Wave 2

For each visit they made to the household, the Wave 2 interviewers responded to eleven questions about the general state of the neighbourhood and on whether they felt safe or unsafe when they visited the household. This information was gathered for both responding and nonresponding households across the UK. Up to 15 visits were made in some cases. In most cases, however, the interviewer gave the same answer regardless of

Table A3.1. Categories of assessment item and scores.

Assessment item	Category	Score
1. How would you rate the general condition of most of the residences or other buildings in the street?	Well kept, good repair and exterior surfaces	0
	Fair condition	1
	Poor condition, peeling paint, broken windows	2
	Badly deteriorated	2
2. Do any of the fronts of residential or commercial units have metal security blinds, gates or iron bars & grilles?	None	0
	Some	1
	Most	2
3. Are there any traffic calming measures in place on the street?	No traffic permitted	0
	Light traffic	0
	Calming + moderate traffic	0
4. How would you rate the volume of traffic on the street?	No calming + moderate	1
	Calming + heavy traffic	1
	No calming +heavy traffic	2
5. Are there any burnt-out cars on the street?	No	0
	Yes	1
6. Is there any of the following: rubbish, litter, broken glass, drug related items, beer cans etc, cigarette ends or discarded packs – in the street or on the pavement?	None or almost none	0
	Yes, some	1
	Yes, just about everywhere you look	2
7. Is there any graffiti on walls or on public spaces like bus shelters, telephone boxes or notice boards?	No	0
	A little	1
	A lot	2
8. Is there dog mess on the pavement?	None	0
	Some	1
	A lot	2
9. Is there any evidence of vandalism such as broken glass from car windows, bus shelters or telephone boxes?	No	0
	Yes	1
10. Are there any adults or teenagers in the street or on the pavements arguing, fighting, drinking or behaving in any kind of hostile or threatening way?	No-one seen in the street or pavement	0
	None observed behaving in hostile ways	0
	Yes, one or two arguing etc.	1
	Yes, at least one group of three or more	2
11. How did you feel parking/walking/waiting at the door in the street?	Very comfortable, can imagine living/working/shopping here	0
	Comfortable – a safe and friendly place	0
	Fairly safe and comfortable	1
	I would be uncomfortable living/working/shopping here	2
	I felt like an outsider, looked on suspiciously	2
	I felt afraid for my personal safety	2

how many times they visited the property and so there was no evidence that interviewers' perceptions changed according to the time of day or day of the week that they were in the area. Consequently, the data used here come from the first visit to each household.

The scoring for the summary score is shown in Table A3.1.

The summary score can vary from zero (positive assessment) to 20 (very negative assessment) (with questions 3 and 4 scored together) but very few scores over 10 were obtained as shown in Table A3.2.

Table A3.2. Distribution of neighbourhood assessment score (n = 16,594).

Score	0	1–3	4–6	7–10	>10
%	34	42	16	6	2

Note Data source: MCS, Wave 2.

7. References

Alberman, E.D. and H. Goldstein. 1970. "The At Risk Register: A Statistical Evaluation." *British Journal of Preventive and Social Medicine* 24: 129–135. Doi: http:/dx.doi.org/10.1136/jech.24.3.129.

Behr, A., E. Bellgardt, and U. Rendtel. 2005. "Extent and Determinants of Panel Attrition in the European Community Household Panel." *European Sociological Review* 21: 489–512. Doi: http:/dx.doi.org/10.1093/esr/jci037.

Blom, A.G. 2014. "Setting Priorities: Spurious Differences in Response Rates." *International Journal of Public Opinion Research* 26: 245–255. Doi: http:/dx.doi.org/10.1093/ijpor/edt023.

Binder, D. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–292. Doi: http:/dx.doi.org/10.2307/1402588.

Box, G.P. and D.R. Cox. 1964. "An Analysis of Transformations." *Journal of Royal Statistical Society, Series B* 26: 211–252.

Browne, W.J. 2009. *MCMC Estimation in MLwiN*. Centre for Multilevel Modelling, University of Bristol. Available at: http://www.bristol.ac.uk/cmm/software/mlwin/features/mcmc.html (accessed July 17, 2017).

Browne, W.J., F. Steele, M. Golalizadeh, and M.J. Green. 2009. "The Use of Simple Reparameterizations to Improve the Efficiency of Markov Chain Monte Carlo Estimation for Multilevel Models with Applications to Discrete Time Survival Models." *Journal of Royal Statistical Society, Series A* 172: 579–598. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00586.x.

Calderwood, L., I. Plewis, S.C. Ketende, and T. Mostafa. 2016. "Evaluating the Immediate and Longer-term Impact of a Refusal Conversion Strategy in a Large Scale Longitudinal Study." *Survey Research Methods* 10: 225–236. Doi: http://dx.doi.org/10.18148/srm/2016.v10i3.6275.

Couper, M. and M. Ofstedal. 2009. "Keeping in Contact with Mobile Sample Members." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 183–203. Chichester: John Wiley. Doi: http://dx.doi.org/10.1002/9780470743874.ch11.

Durrant, G.B. and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from six UK Government Surveys." *Journal of the Royal Statistical Society, Series A* 172: 361–382. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00565.x.

Durrant, G.B., J. D'Arrigo, and F. Steele. 2011. "Using Paradata to Predict Best Times of Contact Conditioning on Household and Interviewer Influences." *Journal of the Royal Statistical Society, Series A* 174: 1029–1049. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00715.x.

Fumagalli, L., H. Laurie, and P. Lynn. 2013. "Experiments with Methods to Reduce Attrition in Longitudinal Surveys." *Journal of the Royal Statistical Society, Series A* 176: 499–519. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01051.x.

Groves, R.M. 2006. "Nonresponse Rates and Non-response Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. Doi: http://dx.doi.org/10.1093/poq/nfl033.

Hawkes, D. and I. Plewis. 2006. "Modelling Non-Response in the National Child Development Study." *Journal of the Royal Statistical Society, Series A* 169: 479–491. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00401.x.

Lynn, P., O. Kaminska, and H. Goldstein. 2014. "Panel Attrition: How Important is Interviewer Continuity?" *Journal of Official Statistics* 30: 443–457. Doi: http://dx.doi.org/10.2478/jos-2014-0028.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R.M. Groves, and T.E. Raghunathan. 2010. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society, Series A* 173: 389–408. Doi: http://dx.doi.org/10.1111/j.1467-985X.2009.00621.x.

Krzanowski, W.J. and D.J. Hand. 2009. *ROC Curves for Continuous Data*. Boca Raton, Fl: Chapman and Hall/CRC. Doi: http://dx.doi.org/10.1201/9781439800225.

Laurie, H. and P. Lynn. 2009. "The Use of Respondent Incentives on Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 205–233. Chichester: John Wiley. Doi: http://dx.doi.org/10.1002/9780470743874.ch12.

Lynn, P. 2017. "From Standardised to Targeted Survey Procedures for Tackling Non-response and Attrition." *Survey Research Methods* 11: 93–103. Doi: http://dx.doi.org/10.18148/srm/2017.v11i1.6734.

McGonagle, K., M. Couper, and R. Schoeni. 2009. "An Experimental Text of a Strategy to Maintain Contact with Families Between Waves of a Panel Study." *Survey Practice* 2(5).

Pepe, M.S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: OUP. Doi: http://dx.doi.org/10.1198/jasa.2005.s19.

Peytchev, A., S. Riley, J. Rosen, and J. Murphy. 2010. "Reduction of Nonresponse Bias in Surveys Through Case Prioritization." *Survey Research Methods* 4: 21–29. Doi: http://dx.doi.org/10.18148/srm/2010.v4i1.3037.

Plewis, I. 2007a. "Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study." *International Journal of Social Research Methodology* 10: 325–334. Doi: http://dx.doi.org/10.1080/13645570701676955.

Plewis, I. (Ed.). 2007b. *The Millennium Cohort Study: Technical Report on Sampling* (4th Ed.). London: Institute of Education, University of London. Available at:

http://www.cls.ioe.ac.uk/library-media/documents/Technical_Report_on_Sampling_4th_Edition.pdf (accessed July 17, 2017).

Plewis, I., S.C. Ketende, H. Joshi, and G. Hughes. 2008. "The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First two Waves of the Millennium Cohort Study." *Journal of Official Statistics* 24: 365–385. Doi: http://dx.doi.org/10.1080/13645570701676955.

Plewis, I., S.C. Ketende, and L. Calderwood. 2012. "Assessing the Accuracy of Response Propensities in Longitudinal Studies." *Survey Methodology* 38: 167–171.

Pudney, S. and N. Watson. 2013. *If at First you Don't Succeed? Fieldwork, Panel Attrition, and Health-Employment Inferences in BHPS and HILDA*, Institute for Social and Economic Research Working Papers, No. 2013–27, ISER, University of Essex. Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2013-27 (accessed July 17, 2017).

Rasbash, J., C. Charlton, W.J. Browne, M. Healy, and B. Cameron. 2009. *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol. Available at: http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html (accessed July 17, 2017).

SAS Institute Inc. 2011. SAS/STAT® 9.3, User's Guide. Cary, NC: SAS Institute Inc. Available at: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm (accessed July 17, 2017).

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators of the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics* 27: 231–253.

Schouten, B. and N. Shlomo. 2017. "Selecting Adaptive Survey Design Strata with Partial R-indicators." *International Statistical Review* 85: 143–163. Doi: http://dx.doi.org/10.1111/insr.12159.

Shlomo, N., C. Skinner, and B. Schouten. 2012. "Estimation of an Indicator of the Representativeness of Survey Response." *Journal of Statistical Planning and Inference* 142: 201–211. Doi: http://dx.doi.org/10.1016/j/jspi.2011.07.008.

Skinner, C. and J. D'Arrigo. 2011. "Inverse Probability Weighting for Clustered Nonresponse." *Biometrika* 98: 953–966. Doi: http://dx.doi.org/10.1093/biomet/asr058.

Watson, N. and M. Wooden. 2009. "Identifying Factors Affecting Longitudinal Survey Response." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 157–182. Chichester: John Wiley. Doi: http://dx.doi.org/10.1002/9780470743874.ch10.

Watson, N. and M. Wooden. 2014. "Re-Engaging with Survey Non-Respondents: Evidence from Three Household Panels." *Journal of the Royal Statistical Society, Series A* 177: 499–523. Doi: http://dx.doi.org/10.1111/rssa.12024.

Received March 2016
Revised June 2017
Accepted June 2017