STATISTICAL GUIDANCE ON OPTIMAL STRATEGIES TO REDUCE NON-RESPONSE IN LONGITUDINAL STUDIES

Ian Plewis and Natalie Shlomo, Social Statistics, University of Manchester

Abstract: We examine strategies for sample maintenance in longitudinal studies which take into account the trade-off between limited resources and ensuring representativeness and the quality of response. We utilize information contained within response propensity models that are widely used in longitudinal research. Based on estimated response propensities, we carry out statistical testing of the effectiveness of 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-off for the propensity to respond when targeting interventions. We use the first four waves of the UK Millennium Cohort Study (MCS) to illustrate these methods. Our results suggest that it is worth reissuing to the field cases that are not located or not contacted at a previous wave but reissuing refusals might not be a good use of resources. The ability to discriminate between respondents and non-respondents is not high and this leads us to question the value of interventions to reduce non-response in longitudinal studies.

Keywords: Longitudinal Studies, Millennium Cohort Study, Response Propensity Models, ROC Curves, Representativity Indicators

6262 Words of text

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 do, however, need to be based on the efficient use of limited resources. As Couper and Ofstedal (2009, p.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 non-located units can be minimized it may be costly to do so." There is, however, little guidance from research that managers of longitudinal studies can draw on to enable them to allocate resources optimally, either within activities such as reducing the number of refusals, or between activities such as tracking on the one hand and increasing cooperation on the other.

This paper aims to fill at least some of this gap in the literature by exploiting information contained within the response propensity models that are widely used in longitudinal research, often to construct non-response weights (Kalton and Flores-Cervantes 2003). Two different but related issues are addressed: (i) how effective are the strategies that are widely and routinely adopted by survey managers to retain sample members in a longitudinal study; (ii) which sample members should be the targets of interventions that have been shown, ideally from experiments, to improve the quality of response.

The paper proceeds as follows. Sample maintenance strategies and their costs and potential benefits are discussed in Section 2. This is followed by Section 3 on response propensity models and two sets of measures derived from them, one set based on the variability of the predicted probabilities of responding and the other on receiver operating characteristic (ROC)

curves. Section 4 presents the study to which the ideas in this paper are applied, the UK Millennium Cohort Study, with illustrative results. The paper concludes in Section 5 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 in the social sciences.

2. Sample maintenance strategies

Suppose that our 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, managers of such studies can choose to assign their limited resources to locating sample members who change address, to reducing non-contact conditional on location, and to increasing cooperation conditional on contact by:

1) Maintaining regular contact with sample members over time through, for example, newsletters and greetings cards, and locating those who move by using administrative records, by collecting stable addresses for respondents at the first wave etc. 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, including the use of incentives to provide information on change of address, in the context of long-running household panel surveys.
2) Minimising non-contact 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 sub-groups of the sample, or by providing incentives to interviewers to attain cooperation from units thought likely to refuse. In their summary of the evidence about the value of incentives, Laurie and Lynn (2009) conclude that they do improve response rates but do not necessarily reduce non-response bias. Moreover, incentives are expensive and suffer from the 'deadweight' problem in that they are given to sample members who would have responded without them. Kaminska et al. (2011) 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 four or five 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 monetary incentives neither improved response rates nor reduced non-response 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.
4) Maintaining a good response rate over time by reissuing to the field at later waves cases that were not productive, for whatever reason, at earlier waves. As Watson and Wooden (2013) 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 reissuing (usually to a different interviewer) cases who did not cooperate initially. Calderwood et al. (2010) describe an experiment built into the fourth wave of the UK Millennium Cohort Study in which this strategy did improve response rates and also reduced non-response bias for some variables. On the other hand, the strategy was expensive as less than a quarter of the reissued cases were converted into productive or partially productive cases.

We can divide these sample maintenance strategies into two broad categories: those that are part of the craft and standard practice of managing longitudinal studies such as tracking mobile households, and those that test specific interventions such as offering incentives to interviewers. Strategies in both categories should, however, be subject to rigorous assessment of their effectiveness.

Raising response rates increases the precision of estimates of change parameters. 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 compared with estimates of change obtained from a longitudinal sample based on a lower response rate. Thus, we would ideally 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) against the costs and benefits of a lower response rate but where this smaller achieved sample more closely mirrors the target population and so non-response bias is also lower. When comparing two strategies, Bethlehem et al. (2011, p. 42) point out that bias is reduced if $Q_2/Q_1 < K_1/K_2$ where Q_i (i=1,2) is the proportion of non-respondents for strategy i (with $Q_2 > Q_1$) and K_i (i=1,2) is the difference between the estimates in the respondent and non-respondent strata ($K_2 < K_1$). Consequently we need:

i. Evidence about the efficacy of different strategies to reduce the number of cases not located and not contacted and to increase levels of cooperation, in terms of increasing response and reducing bias.

ii. Information about the actual costs of implementing these strategies.

iii. An assessment of the opportunity costs of each strategy which, in turn, requires a valuation to be put either on a reduction in mean square error (MSE) or separately on 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.

3. Response propensity models

There are many instances in the literature of studies that have modelled the predictors of non-response in longitudinal surveys mostly to generate non-response weights: for example, Behr et al. (2005); Hawkes and Plewis (2006); Watson and Wooden (2009) and, for the Millennium Cohort Study, 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, productive vs. non-productive - is linked to a set of explanatory variables, using either a logit or probit link (or their multivariate equivalents). A simple example is:

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

where $\rho_i = E(r_i)$ is the probability of not responding for unit i (i = 1..n); $r_i = 0$ for a response and 1 for non-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 not responding $\hat{\rho}_i$ where

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

These predicted probabilities of not responding can be used in a number of ways. They can be ordered and plotted against their ranks in what Huang et al. (2009) describe as a predictiveness curve; their standard deviations can be used to construct R (for representativity) indicators as shown by Schouten et al. (2009) and elaborated in the next section; and 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 here to examine the most efficient way of targeting interventions.

There are different ways of specifying models like (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. This is explored further in the context of our example.

3.1 Representativity indicators

Schouten et al. (2009; 2011) have developed quality indicators for measuring the extent to which a survey is representative of the population under investigation, i.e. the degree to which respondents and non-respondents in a survey differ from each other. The Representativity indicators (R-indicators) are based on the variability in the response propensities for a sample *s* drawn from a population *U*. The R-indicator is estimated by:

$$\hat{R}_\rho = 1 - 2\hat{S}_\rho \quad (3)$$

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), $\hat{\bar{\rho}}_U = (\sum_s d_i\hat{\rho}_i)/N$ 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 not to respond for each case is equal to the overall non-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 response propensity model. $\hat{R}_\rho = 0$ 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 3 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 particular categorical variables of interest and for the 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:

$$\hat{R}_{\rho(u)} = \hat{S}_B(\hat{\rho} \mid Z) \qquad (4)$$

where $\hat{S}_B^2(\hat{\rho} \mid Z) = \sum_{k=1}^{K} \frac{\hat{N}_k}{N} (\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U)^2$, $\hat{\bar{\rho}}_k$ 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:

$$\hat{R}_{\rho(u),k} = \hat{S}_B(\hat{\rho} \mid Z = k) \frac{\left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right)}{\left|\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right|} = \sqrt{\frac{\hat{N}_k}{N}} \left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right) \tag{5}$$

Note that $\hat{R}_{\rho(u),k}$ can be positive (under -representation) or negative (over-representation). Conditional partial R- indicators measure the remaining variance due to variable *Z* within sub-groups formed by all other remaining variables, denoted by X^-. In other words, the conditional partial indicators tell us whether a variable *Z* contributes to the explanation of non-response conditional on the other variables in the model. The conditional partial indicators are not considered further in this paper.

All the published applications of R-indicators have been to cross-sectional surveys as a means of assessing the bias-reducing value of additional callbacks and other strategies to increase the achieved sample size, 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. It is, however, possible to exploit the more detailed information available from the first wave of a longitudinal design to assess the representativeness of later waves in terms of the wave one sample and hence as an aid to judge the value of the sample maintenance strategies described earlier.

In the application shown in Section 4 based on the Millennium Cohort Study, the design weights are calculated relative to the sample in wave 1 and represent the disproportionate sampling within strata. Therefore, in the calculation of the formula for the R-indicator in (3) and unconditional R-indicators in (4) and (5), 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 $\hat{\bar{\rho}}_k = \frac{1}{\hat{n}_k} \sum_{s_k} d_i \hat{\rho}_i$.

3.2 Receiver Operating Characteristic curves

A widely used method of assessing the accuracy of models for binary or categorical outcomes is to estimate their goodness-of-fit by using one of several possible pseudo-R^2 statistics. Apart from their rather arbitrary nature, which thus makes comparisons across datasets difficult, estimates of pseudo-R^2 are not especially useful in this context because they assess the overall fit of the model rather than distinguish between its accuracy for discriminating between non-respondents (the true positive rate) and respondents (the true negative rate) separately. Consequently, we use measures based on ROC curves i.e. the plot of the sensitivity or true positive rate against the false positive rate.

The area enclosed by the ROC curve and the x-axis, known as the AUC (area under the curve), is of particular interest and this can vary from 1 (when the model for predicting non-response perfectly discriminates between respondents and non-respondents) 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 non-respondent, 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. The Youden index, defined as the maximum vertical distance from the diagonal to the ROC curve, is an alternative summary of accuracy. See Krzanowski and Hand (2009) for a detailed discussion of how to estimate ROC curves and measures derived from them.

The ROC represents the balance between the true and false positive rates (TPF and FPF) for different cut points on the response propensity scale (i.e. the $\hat{\rho}_i$) . The slope of the ROC, *a(v)*, is the ratio of two conditional densities:

$a(v) = f(v \mid r = 1) / f(v \mid r = 0)$ where v is the estimated linear predictor from the response propensity model (1).

Given a desire to intervene to prevent non-response, we then want to target our resources in the most efficient way. We can do this by intervening so that everyone with a response propensity above an optimum cut point is eligible to receive the intervention and nobody with a response propensity below the cut point receives it. One way of determining the optimum cut point is to minimise a cost function such as the one set out by Pepe (2003, p.32) for the overall cost of non-response per case (*TC*):

$$TC = C_{NR}^{I+} TPF \hat{\bar{\rho}}_s + C_{NR}^{I-} (1 - TPF) \hat{\bar{\rho}}_s + C_R^{I+} FPF (1 - \hat{\bar{\rho}}_s) \quad (6)$$

where the first cost term on the right-hand side of (6) is the actual cost of intervening (*I+*) when the case is a non-respondent (*NR*), and the second and third cost terms are misclassification costs; $\hat{\bar{\rho}}_s$ is the prevalence of non-response. Then the optimal cut point, a^o, is determined from *TC* by minimising *TC* with respect to *TPF* for a fixed *FPF*, i.e. the slope of the ROC curve is:

$$a^o = O * F \quad (7)$$

and O is the odds of being a respondent (and therefore O is usually greater than one), 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 opportunity cost of failing to intervene for a non-respondent (the false negatives) minus the actual cost of intervening when the prediction to be a non-respondent is correct (the true positives), i.e. $F = C_R^{I+}/(C_{NR}^{I-} - C_{NR}^{I+})$ and the denominator is assumed to be positive (Pepe 2003, p.72). Alternative optima appear in the literature. Krzanowski and Hand (2009 p.24) focus solely on the costs of misclassification and so F in (6) is then C_R^{I+}/C_{NR}^{I-} with O remaining unchanged. Other authors (e.g. Fluss et al. 2005) argue that the optimum threshold is the one determined by the Youden index. This is independent of the prevalence of non-response and the costs of misclassification. We return to these issues in the context of our example in Section 4.

Ideally, any decisions about how to intervene at wave t would be based on a response propensity model for that wave but, of course, the required information on 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 assume 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.

4. Millennium Cohort Study

The wave one sample of the UK Millennium Cohort Study (MCS) includes 18,818 babies in 18,552 families born in the UK over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards 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 initial response rate was 72%. 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 design weights vary from 2.0 (England advantaged stratum) to 0.23 (Wales disadvantaged stratum). The first four waves took place when the cohort members were (approximately) nine months, 3, 5 and 7 years old. 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.

It has been standard practice in MCS to reissue all eligible cases at wave t, conditional on their being in the observed sample at wave one. Cases become ineligible by virtue of emigration, being in institutional care or child death. There are some exceptions to this practice: so-called 'hard' refusals were never reissued, and the majority of eligible cases that were unproductive at both waves two and three were not reissued at wave four.

4.1 Illustrative results from MCS

In order to assess robustness to different specifications, we consider three response propensity models that are used to predict response behaviour after wave one and which are increasingly complex in terms of their explanatory variables. The first model is similar to the one used in Plewis (2007a), the explanatory (or auxiliary) variables are all measures obtained at wave one of MCS and are listed in Appendix 1. The second includes an additional variable that became available from survey managers after wave one and is described in Plewis et al. (2008): whether the main respondent changed address between waves one and two and the interactions of this variable with tenure and type of accommodation. The first two models allow for the sample design in terms of its disproportionate stratification and clustering using *svy* commands in STATA. The third model , however, 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 essentially the same as that used in a related context by Durrant and Steele (2009):

$$\text{logit}(\rho_{ij}) = \sum_{k=1}^{K} \beta_k x_{kij} \quad + \sum_{p=1}^{P} \gamma_p z_{pj} + u_j \qquad (8)$$

where:

ρ_{ij} is the probability of not responding for respondent *i* ($i = i..n_j$) in cluster (i.e. electoral ward) *j* ($j = 1..J$).

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.

The model was estimated using Markov chain Monte Carlo (MCMC) methods available in the *MLwiN* software (Rasbash et al. 2009; Browne 2009), based on 40K iterations following a burn-in of 5K, with non-informative priors throughout and supplemented by orthogonal parameterisation and parameter expansion as described by Browne et al. (2009) to improve convergence.

The predicted values $\hat{\rho}_{ij}$ from equation (8) include the Bayes estimates $\hat{u}_j$ - the means of the 40K MCMC iterations - estimating the deviation from zero of the proportion of non-response for each cluster. Hence:

$$\hat{\rho}_{ij=} \exp(\sum_{k=1}^{K} b_k x_{kij} + \sum_{p=1}^{P} c_p z_{pj} + \hat{u}_j)/[1 + \exp(\sum_{k=1}^{K} b_k x_{kij} + \sum_{p=1}^{P} c_p z_{pj} + \hat{u}_j)] \quad (9)$$

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 non-response is cluster-specific nonignorable (i.e. where non-response 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 non-respondents 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 in the MCS data, accounting for 322 cases.

4.2 Representativeness

R-indicators are used here to provide evidence about the utility of different reissuing strategies. These reissuing strategies are labelled (i) S (for standard practice) so S2, S3 and S4 refer to the standard practices at waves two, three and four in MCS; (ii) P, the hypothetical strategy of only reissuing productive cases from previous waves so P3.2 refers to a strategy of only reissuing at wave three cases that were productive at wave two, P4.23 refers to only reissuing at wave four cases that were productive at waves two and three, P4.3 to only reissuing at wave four cases that were productive at wave three (including some that were not productive at wave two); (iii) C (for cooperation), the hypothetical strategy of not reissuing refusals from previous waves so C3.2 refers to a strategy of not reissuing at wave three refusals from wave two, C4.23 refers to not reissuing at wave four refusals from waves two and three, C4.3 to not reissuing refusals just from wave three; (iv) W4 the hypothetical

strategy of not reissuing at wave four cases that were not productive at wave two but were productive at wave three – sometimes known as wave non-respondents.

Results (not shown here) show that the estimates of the R-indicators and partial R-indicators are very similar across the three models listed in section 4.1 and hence we show results for model 2 only. 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, weights and stratification. Appendix 3 shows how to derive the standard errors of the R-indicators under the complex survey design.

Figure 1 gives the results of the reissuing strategies for the R-indicators along with their confidence intervals. The sample sizes for each model after omitting ineligible cases are 18,148, 17,990 and 17,819 for waves two to four. Note that the size of the productive sample at wave one is 18,552 but a few cases were omitted from the response propensity models because of item non-response at wave one.

When making comparisons of the reissuing strategies, it's important to keep in mind two caveats:

(1) Non-overlapping confidence intervals do not necessarily mean statistical significance but do provide some indication of significance if the sample sizes are nearly equal ;

(2) We are in fact carrying out multiple comparisons of reissuing strategies so the significance level for the confidence intervals in the figures have been changed according to the Bonferroni correction, e.g. in Figure 1, the confidence interval is at the 0.05/(10 strategies) = 0.005 significance level. When comparing only two or three reissuing strategies, the confidence intervals should be narrower.

Figure 1: Estimates of R-indicators (with confidence intervals adjusted for multiple comparisons) for reissuing strategy according to model 2

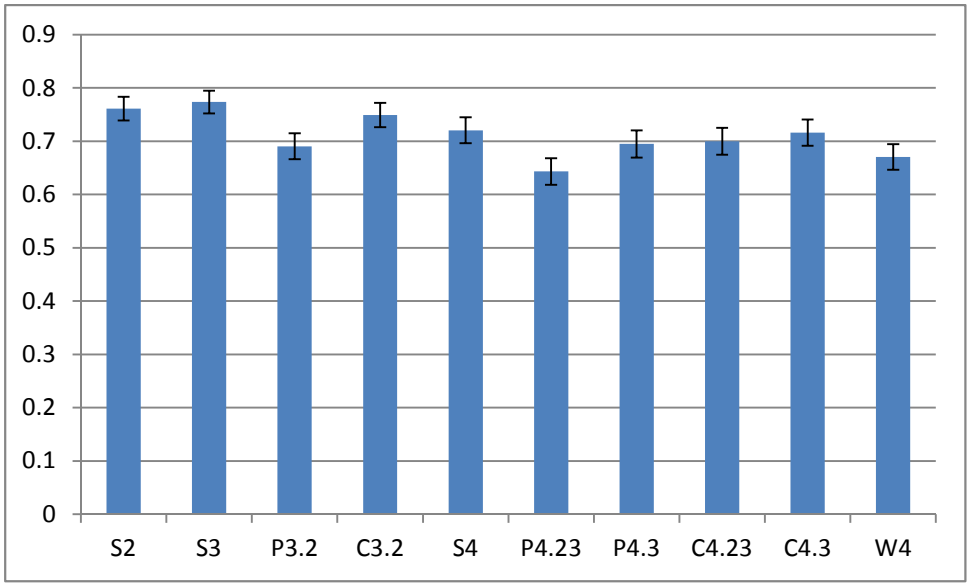


Table 1: Number of cases lost and the percentage of actual productive sample at each wave according to reissuing strategy

Row label	Cases lost	Percentage of actual productive sample
S2	n.a.	-
S3	n.a.	-.
P3.2	1,444	9.5%
C3.2	473	3.1%
S4	n.a.	-
P4.23	1,668	12.0%
P4.3	639	4.6%
C4.23	536	3.9%
C4.3	216	1.6%
W4	1,029	7.4%

The two columns labelled S2 and S3 show little difference in representativeness between waves two and three of the MCS sample given standard reissuing practice. On the other

hand, the column labelled S4 shows lower representativeness at wave 4 compared to the first two waves using standard reissuing practice. Columns labelled P show that representativeness falls if only productive cases from previous waves are reissued and the estimates are all lower than for S3; the estimates for P4.3 are closer to S4 than those for P4.23 are.

Comparing C3 to S3 suggests that representativeness is less compromised if refusals from previous waves are not reissued since the R-indicators are nearly equal. Comparing the strategies of cases that refused just at wave three not being reissued at wave four (C4.3) to those that refused at either wave two or wave three (C4.23), we see a slight increase in representativeness for C4.3 but both strategies have similar representativeness to S4 with confidence intervals that do not overlap. Finally, the column labelled W4 shows that representativeness is reduced if wave non-respondents are not reissued. The number of cases that would have been lost and the percentage of actual productive sample at each wave, all other things being equal, as a result of the different reissuing policies is shown in Table 1.

Figures 2a and 2b gives estimates of unconditional partial R-indicators at the variable level 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. The first thing to be noticed is that the unconditional partial R-indicators for the two variables across the reissuing strategies are all significantly different from zero, even when showing confidence intervals under the full multiple comparison correction. Thus, the distributions of highest educational qualifications and, to a lesser extent, ethnic group are biased at waves two, three and four when compared with the

distributions at wave one. Figures 2a and 2b also show that only issuing productive cases from waves two and three at wave four (i.e. P4.23 in Figure 1) leads to poorer representativeness for both variables. Generally, the results in Figures 2a and 2b are in line with those from Figure 1.

Figure 2a: Estimates of unconditional partial R-indicators (with confidence intervals adjusted for multiple comparisons) for main respondent's highest educational qualifications according to reissuing strategy (model 2)

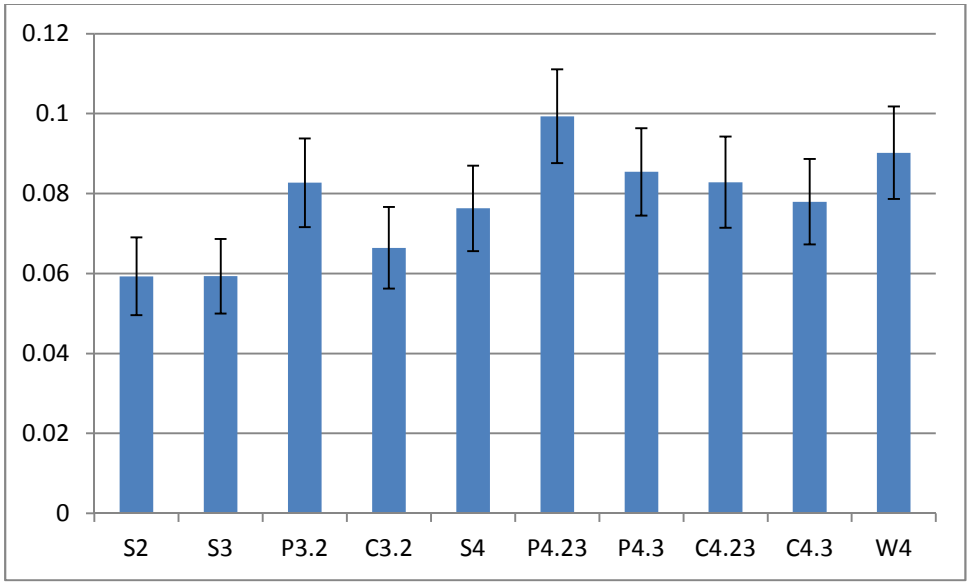
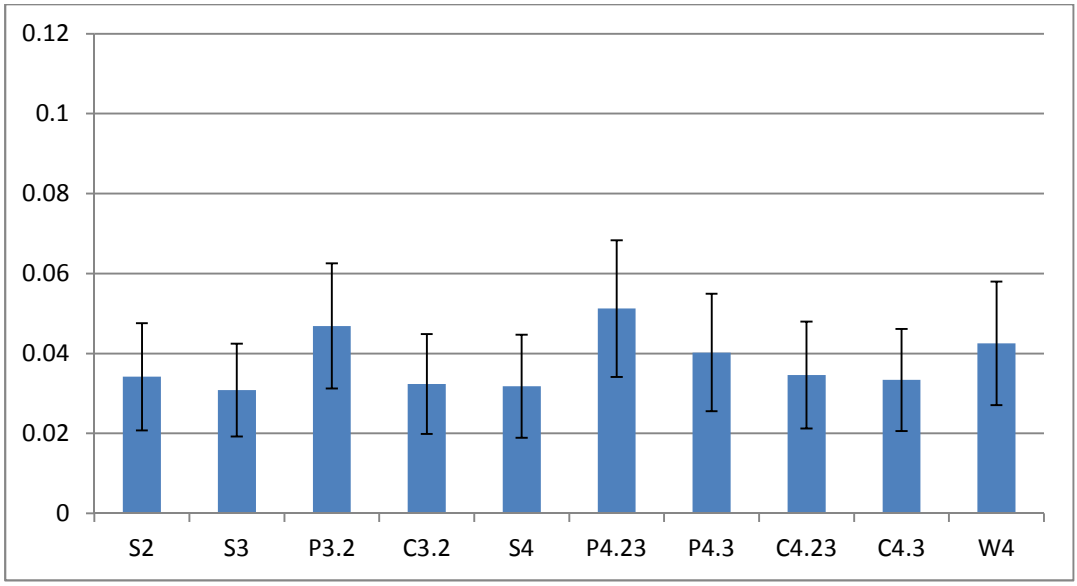


Figure 2b: Estimates of unconditional partial R-indicators (with confidence intervals adjusted for multiple comparisons) for ethnic group according to reissuing strategy (model 2)



The estimates for individual categories (not presented here) indicate over-representation of those with higher rather than lower or no qualifications, and of white British compared with minority ethnic groups.

We can use the information from the partial R-indicator estimates to reassess the reissuing strategy. For example, we might decide not to reissue the cases that were unproductive at earlier waves (column P4.23) only if they belonged to the majority ethnic group or who had some educational qualifications. The estimate of the R-indicator then increases from 0.65 to 0.71 for model 2 with 902 cases lost and so the strategy is as effective as not reissuing unproductive cases just from wave three (column P4.3) and, in fact, is essentially the same as the standard practice (S4).

4.3 ROC

Table 2 shows how discrimination between respondents and all non-respondents varies by model and wave. In all cases, the AUC were estimated assuming a binormal model (Pepe

2003), and the Gini coefficients and their standard errors were derived from these estimates. We see that discrimination at wave two is slightly improved when allowing for the sample design and only slightly reduced at waves three and four even though the models are based solely on wave one variables.

Table 2: Gini Estimates (s.e.) by Model and Wave

	Model 1	**Model 2**	**Model 3**
Wave 2	0.39 (0.0100)	0.39 (0.0100)	0.42 (0.0095)
Wave 3	0.37 (0.0099)	0.37 (0.0099)	0.36 (0.0100)
Wave 4	0.37 (0.0090)	0.37 (0.0090)	0.39 (0.0088)

Note: Sample sizes as for Figure 1.

We now consider how we might target interventions designed to prevent refusals or not being located at wave t, conditional on being in the wave one sample. Here we focus on wave three when some data was obtained from 80% of the eligible wave one sample and refusals were nearly twice as common as other kinds of unproductive cases. Our interest is in directing interventions already known to have an effect on converting refusals or not located into productive cases, i.e. targeting cases most likely to benefit in terms of their estimated propensity to respond. We also consider how robust the targeting is to changes in the response propensity model and to misclassification of the outcome variable. 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 non-contacts at wave two and refusals at wave three as non-contact can sometimes be a hidden refusal. We also extend model two for refusals by including an assessment of the neighbourhood by interviewers at the times they called at sample households at wave two. This variable, which can be thought of as paradata in Kreuter et al.'s (2010) terms, was not available for the 'not located' group. It is described

in more detail in Appendix 2. Table 3 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 three for the more encompassing definition of refusal. Discrimination is higher for not located than it is for refusal.

Table 3: Gini Estimates (s.e.) by Model, Wave 2

Type of non-response (Prevalence, wave two)	**Model 1**	**Model 2**	**Model 3**
Refusal (i) (10.0%)	0.35 (0.015)	0.40 (0.017)	0.52 (0.016)
Refusal (ii) (11.0%)	0.35 (0.014)	0.41 (0.016)	0.43 (0.016)
Not located (4.2%)	0.49 (0.026)	n.a.	0.54 (0.022)

Note: Sample sizes: (a) refusal, model 1 – 16468 and 16627; (b) refusal, models 2 and 3 – 15647 and 15781; (c) not located, models 1 and 3 – 15403.

The estimates of the Youden index show a similar pattern to the Gini estimates across models and types of non-response but when they are used to estimate an optimum threshold the corresponding response propensities are small and this leads to potential intervention groups that are unrealistically large (n > 5000). Consequently, we focus on estimating optima that take account of prevalence and costs as set out in Section 3.2. The estimates of O from equation (7) for refusal as defined in (i) above are 8.6 (model 1) and 10.5 (models 2 and 3) and we consider three values of the cost ratio F: 0.33, 0.8 and 1.5. It is, in fact, not possible to estimate a 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 values of $\hat{\rho}_i$ are 0.28, 0.32 and 0.23; for $F = 0.8$ for model 3, the cut-off for $\hat{\rho}_i$ is 0.46 (these estimates were obtained after applying Box-Cox transformations to the conditional distributions).

Table 4 shows how the optimal 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 804 (F = 0.33, model 3) down to 82 (F = 0.8, model 3).Table 4 also shows what proportions of the intervention groups actually refused at wave two and the response outcomes for these groups at wave three. The final row shows how much larger the sample at wave three 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.

Table 4: Intervention Groups 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
Refusal + non-contact, wave 3 (%)	34	33	30	40
Maximum increase in wave 3 sample (n)	155	73	201	28

Very similar results to those shown in Table 4 are obtained when the more inclusive definition of refusal (i.e. (ii) above) is used.

It is only possible to estimate a cut-off for not located when F = 0.33 for model 3, corresponding to $\hat{\rho}_i = 0.29$ and an intervention group of size 33, a third of which were not located at wave two and 18% at wave three.

5. Conclusions

The results in the previous section are only illustrative and they are confined 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. Nevertheless, they do suggest that the practice of reissuing refusals from wave t at a subsequent wave (or waves) might not be cost-effective, given that excluding them leads to only a small loss in representativeness, that the proportion of cases lost is also relatively small, and bearing in mind that many of the reissued refusals continue to refuse at subsequent waves. The results in Figure 1 do, however, imply that it is worth reissuing cases that were not located or not contacted (with the majority falling into the not located group). These conclusions appear to be robust to changes in the specification of the response propensity model although only one model was shown. 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 one also maintain representativeness with respect to the target population.

The main point to emerge from the ROC analyses is to question the value of implementing intervention strategies to reduce non-response in longitudinal studies. There are a number of reasons for this. The accuracy of response propensity models is not high and hence they do not clearly discriminate between productive and different kinds of unproductive cases. This means that the ROC curves are relatively shallow. Given that the prevalence of types of non-response at any one wave is usually relatively low (and so O 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 opportunity cost of failing to intervene to prevent a case of non-response is three times the cost of intervening unnecessarily (i.e. $F = 0.33$) is arguably optimistic 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 not be cost-effective to intervene to reduce the size of the not located group and only cost-effective to have an intervention directed at a very small group to prevent refusal. 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 non-response. The results in Table 2 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. Other decision rules, not based on ROC curves, are also possible. For example, 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 paper has indicated how to use response propensity models in a way that can help 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.

References

Alberman, E.D. and Goldstein, H. (1970), " The At Risk Register: A Statistical Evaluation," *British Journal of Preventive and Social Medicine*, 24,129-135.

Behr, A., Bellgardt, E. and Rendtel, U. (2005), "Extent and Determinants of Panel Attrition in the European Community Household Panel," *European Sociological Review,* 21, 489-512.

Bethlehem, J., Cobben, F. and Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. New Jersey: Wiley and Sons.

Binder, D. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279-292.

Browne, W. J. (2009), *MCMC Estimation in MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Browne, W. J., Steele, F., Golalizadeh, M., and Green, M.J. (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, 579-98.

Calderwood, L., Plewis, I., Ketende S.C. and Taylor, R. (2010), "Experimental Testing of Refusal Conversion Strategies in a Large-scale Longitudinal Study," *CLS Working Paper 2010/9*. London: Centre for Longitudinal Studies.

Couper, M. and Ofstedal, M. (2009), "Keeping in Contact with Mobile Sample Members," in *Methodology of Longitudinal Surveys* (edited by P. Lynn), 183-203, Chichester: John Wiley.

Durrant, G. B. and Steele, F. (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, 361-82.

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

Fluss, R., Faraggi, D. and Reiser, B. (2005), "Estimation of the Youden Index and its Associated Cutoff Point," *Biometrical Journal*, 47, 458-472.

Fumagalli, L., Laurie, H. and Lynn, P. (2013), "Experiments with Methods to Reduce Attrition in Longitudinal Surveys," *Journal of the Royal Statistical Society, Series A,* 176, 499-519.

Groves, R. M. (2006), " Nonresponse Rates and Non-response Bias in Household Surveys," *Public Opinion Quarterly* 70, 646-75.

Hawkes, D. and Plewis, I. (2006), "Modelling Non-response in the National Child Development Study," *Journal of the Royal Statistical Society, Series A*, 169, 479–91.

Huang, Y. and Pepe, M. S. (2009), "A parametric ROC Model-based Approach for Evaluating the Predictiveness of Continuous Markers in Case-control Studies," *Biometrics*, 65, 1133-44.

Kalton, G. and Flores-Cervantes, I. (2003), " Weighting methods," *Journal of Official Statistics*, 19, 81-97.

Kaminska, O., Lynn, P. and Goldstein, H. (2011), "*Panel Attrition: How Important is it to Keep the Same Interviewer?"* Institute for Social and Economic Research Working Papers, No. 2011-02.

Kreuter, F., Olson, k., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010), "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys," *Journal of the Royal Statistical Society, Series A*, 173, 389-408.

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

Laurie, H. and Lynn, P. (2009), "The Use of Respondent Incentives on Longitudinal Surveys," in *Methodology of Longitudinal Surveys* (edited by P. Lynn), 205-233, Chichester: John Wiley.

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

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: OUP.

Peytchev, A., Riley, S., Rosen, J. and Murphy, J. (2010), "Reduction of Nonresponse Bias in Surveys Through Case Prioritization," *Survey Research Methods* 4, 21-29.

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.

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

Plewis, I., Ketende, S.C., Joshi, H., and Hughes, G. (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.

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

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009), *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

SAS Institute Inc. 2011. SAS/STAT® 9.3, User's Guide. Cary, NC: SAS Institute Inc.

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

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

Shlomo, N., Skinner, C. and Schouten, B. (2012), "Estimation of an Indicator of the Representativeness of Survey Response," *Journal of Statistical Planning and Inference*, 142, 201-211.

Skinner, C. and D'Arrigo, J. (2011), "Inverse Probability Weighting for Clustered Nonresponse," *Biometrika*, 98, 953-966.

Watson, N. and Wooden, M.(2009), "Identifying Factors Affecting Longitudinal Survey Response," in *Methodology of Longitudinal Surveys* (edited by P. Lynn), 157-182, Chichester: John Wiley.

Watson, N. and Wooden, M. (2013), "Re-engaging with survey non-respondents: evidence from three household panels," *Journal of the Royal Statistical Society, Series A*, forthcoming.

Appendix 1

Predictors of non-response

1. Family income (6 ordered categories: 1.8%; 26%; 33%; 20%; 14%; 5.0%)
2. Ethnic group of cohort child (White British (83%); Mixed (3.0%); Indian (2.5%); Pakistani/Bangladeshi (6.9%); Black/Black British (3.6%); Other (1.4%))
3. Accommodation type (House (85%); other (15%))
4. Tenure (Own (58%); rent (36%); other (6.4%))
5. Main respondent's age (< 30 (50%); 30+ (50%))
6. Main respondent's educational qualifications (None (20%); NVQ1-5 (3.3%; 12%; 8.4%; 9.3%; 44%); other/overseas (2.8%))
7. Cohort child breast fed (Yes: 67%)
8. Longstanding illness, main respondent (Yes: 21%)
9. Parental status (2 (83%) or 1 parent family (17%))
10. Main respondent voted in last general election (Yes: 51%)
11. Gave consent to record linkage (Yes: 93%)
12. Provided a stable address at wave one (Yes: 82%)

Appendix 2

Interviewer assessments of the neighbourhood, MCS wave 2.

For each visit they made to the household, the wave two interviewers responded to 11 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 non-responding 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 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 as follows:

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	No calming+ moderate	1
	Calming + heavy traffic	1

street?	No calming +heavy	2
5. Are there any burnt-out cars on the street?	No Yes	0 2
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 Yes, some Yes, just about everywhere you look	0 1 2
7. Is there any graffiti on walls or on public spaces like bus shelters, telephone boxes or notice boards?	No A little A lot	0 1 2
8. Is there dog mess on the pavement?	None Some A lot	0 1 2
9. Is there any evidence of vandalism such as broken glass from car windows, bus shelters or telephone boxes?	No Yes	0 2
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 None observed behaving in hostile ways Yes, one or two arguing etc. Yes, at least one group of three or more	0 0 1 2

11. How did you feel parking/walking /waiting at the door in the street?	Very comfortable, can imagine living/ working/shopping here Comfortable - a safe and friendly place Fairly safe and comfortable I would be uncomfortable living/working/shopping here I felt like an outsider, looked on suspiciously I felt afraid for my personal safety	0 0 1 2 2 2

The summary score can vary from zero to 20 but very few scores over 10 were obtained as shown in Table A2.1.

Table A2.1: Distribution of neighbourhood assessment score (n = 16594)

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

Appendix 3

Calculation of 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:

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

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 SurveyLogistic procedure (SAS Institute Inc. 2011). The second term therefore is calculated as in Shlomo et al. (2012) but replacing the covariance term presented under simple random sampling with the sandwich estimator.

Following Shlomo et al. (2012), for the first term we estimate $var_s\left[\sum_{i\in s} u_i\right]$ where u_i is replaced by $d_i\left(\hat{\rho}_i - \hat{\bar{\rho}}_U\right)^2$ according to the complex survey design. Let n_h be the size of stratum *h*, then the variance is estimated as:

$$var_s\left(\sum_{i\in s} u_i\right) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (u_{hi.} - u_{h..})^2 \qquad \text{(A1)}$$

where f_h is the sample fraction in stratum *h*, $u_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij} u_{hij}$ where w_{hij} is the survey weight of unit j in in cluster *i* and stratum *h,* and m_{hi} is the number of units in cluster *i* and stratum *h*, and $\bar{u}_{h..} = \left(\sum_{i=1}^{n_h} u_{hi.}\right) / n_h$.

To obtain the estimated variance of the unconditional partial R-indicator at the variable level Z as shown in Figures 2a and 2b: $\hat{S}_B(\hat{\rho} \mid Z) = \sqrt{\sum_{k=1}^{K} \frac{\hat{N}_k}{N} (\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U)^2}$, we note that $\hat{S}_B^2(\hat{\rho} \mid Z)$ 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*.