A modified version of this manuscript has been accepted by the Journal of the Royal Statistical Society, Series A, 12 April 2016. Please refer to this paper for citation.

To cite:

Vassallo, Rebecca; Durrant, Gabriele B.; and Smith, Peter W.F. (2016) Separating Interviewer and Area Effects Using a Cross-Classified Multilevel Logistic Model: Simulation Findings and Implications for Survey Designs, *Journal of the Royal Statistical Society*, *Series A*, *forthcoming*.

Separating Interviewer and Area Effects Using a Cross-Classified Multilevel Logistic Model: Implications for Survey Design

Vassallo, Rebecca[1]$_I$; Durrant, Gabriele[1]; and Smith, Peter[1]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.

$_I$*E-mail: rebvas@gmail.com*

Abstract

This work is motivated by an application to separate area and interviewer effects on survey nonresponse which are often confounded. The study aims to provide practical recommendations for future study designs by identifying the smallest total sample sizes and the most geographically-restrictive and cost-effective interviewer allocations required to adequately distinguish between the interviewer and area effects. It is unclear how much interpenetration is needed for a cross-classified multilevel model to work well and to reliably estimate the two higher-level effects. This paper investigates the properties of cross-classified multilevel models under various survey conditions.

Key words: cross-classification, interpenetration, interviewer effects, area effects, multilevel models.

1. Introduction

In face-to-face surveys researchers are often interested in the effect of interviewers on survey estimates or survey processes. One such relationship of interest is the effect of interviewers on nonresponse (Blom et al., 2010; Durrant & Steele, 2009; Durrant et al., 2010; Campanelli & O'Muircheartaigh, 1999; Pickery & Loosveldt, 2002; Pickery et al., 2001; Haunberger, 2010). Since interviewers usually work in a restricted geographical area any interviewer effect identified could simply reflect area differences in the geographic propensity to cooperate in survey requests. Therefore, a particular estimation problem pertains to the identifiability of area and interviewer variation. In a random experiment an interpenetrated sample design would be employed, where each sampled case is allocated randomly to interviewers irrespective of their area. This is considered the gold standard for separating interviewer effects from area effects for face-to-face surveys, but is not implemented in survey practice owing to restrictions in field administration capabilities and survey costs (Schnell & Kreuter, 2005; Campanelli, & O'Muircheartaigh, 1999). A compromise which is achievable in a real survey setting is partial interpenetration. Partial interpenetration exists where interviewers are not fully nested within areas, as one interviewer may work in more than one area, and sampling cases in one area may be designated to more than one interviewer. In the case of partial interpenetration a cross-classified multilevel model specification which considers both interviewer and area random terms has been suggested to distinguish between the two sources of variation (Von Sanden, 2004). Although a range of papers have used such models to distinguish between area and interviewer effects (Campanelli & O'Muircheartaigh, 1999; Durrant et al., 2010; Schnell & Kreuter, 2005), it is unclear how much interpenetration may be needed for a cross-classified multilevel model to work well and to reliably estimate interviewer and area effects. In particular, in circumstances of small sample sizes and low degrees of interpenetration in the dispersion of interviewers across areas, problems of biased estimates and low power for significance tests may arise. Some previous studies (Maas & Hox, 2005; Moineddin et al., 2007; Paccagnella, 2011; Rodriguez & Goldman, 1995; Theall et al., 2011) have looked at the properties of estimators and the power of significance tests for two-level models. However, questions regarding how well cross-classified multilevel model parameters can be estimated have not yet been explored.

This study examines the implications of various practical limitations in the assignment of cases from different areas to interviewers within a range of scenarios through a simulation study. These different scenarios include different total sample sizes, group sizes (interviewer caseload), number of groups (number of interviewers), overall rates of response, and the percentage variance attributable to area and interviewer effects. Interviewer-area classifications are restricted to possible interviewer work allocations, and selected values for the other factors represent realistic values, making the simulation results relevant to survey practice. The implications are assessed in terms of bias, standard error, confidence interval coverage, correlation of the two variance estimators and power of significance tests. The study also examines the smallest interviewer pool and the most geographically-restrictive and cost-effective interviewer case allocation required for acceptable levels of bias and power for typical survey scenarios. By suggesting minimal sample sizes and interviewer dispersal patterns to guide survey design and administration, and by shedding light on the accuracy and precision of the estimates and the power of their tests of significance in multilevel modelling, this study contributes to different areas of research: study design and parameter estimation (Paccagnella, 2011).

Although the factor conditions and the application considered here are specific to survey design and the exploration of interviewer effects on nonresponse, the same problem of identifiability may arise in other settings. For example, other survey design applications may consider the variation in the response to questionnaire items attributable to interviewers, with the aim of quantifying any interviewer influence on responses (measurement error). Other applications with similar design issues can also be envisaged. For example, health studies may be investigating the influence of community physiotherapists in the rehabilitation of patients having undergone orthopaedic surgery. While each patient is associated with their respective physiotherapist, the hospital at which the surgery was undergone must also be taken into account in evaluating their health outcome. Travelling distances and monetary restrictions will mean that individual physiotherapists are assigned home visits to patients within the same local health authority, which matches a specific hospital. Within practical limitations, with a greater geographical spread of cases allocated to each physiotherapist, each physiotherapist will be treating patients from different hospitals, allowing for accurate estimates of the effect of the post-op services on

rehabilitation to be produced. This study can shed light on the amount of cross-classification between hospitals and physiotherapists required for adequate estimates.

The remainder of the paper is structured as follows. First a review of multilevel models and their mathematical properties are presented, followed by an explanation of the use of multilevel models for the analysis of interviewer effects on nonresponse. Then a review of previous work exploring the properties of cross-classified and two-level hierarchical models is given. Section 3 presents the details of the design of the simulation study and the analysis carried out. Next results are presented, followed by a discussion and conclusions.

2. The Multilevel Cross-Classified Model

2.1. Model Specification

The independent errors assumption in standard regression analysis is often not valid for social science data. Individual observations which pertain to some kind of common higher-level grouping – such as school, family, neighbourhood or work organisation – may have similarities arising from the common context which give rise to dependency amongst their observations. Multilevel modelling allows for an extension of the error term included in standard regression analysis to be able to adjust for such dependencies (Goldstein, 2011). Consequently, multilevel models allow the variation in the outcome variable to be partitioned into various sources, these being both individual and group sources. Group similarities are considered as substantively interesting rather than as a model assumption infringement which needs to be accounted for, thus allowing the exploration of significant individual and group influences as well as any possible interactions between these two factors on the individual-level outcome of interest. As well as allowing for a detailed analysis of predictors defined at the cluster-level (also called contextual effects) through the inclusion of a higher-level random effect and contextual or aggregate fixed effects, multilevel models allow for the inclusion of data at the individual level. This helps avoid loss of information at the individual level, a smaller sample size, and the risk of ecological fallacy from analysing aggregated data. Such models do not assume that all contextual effects are included through observable predictors as in a contextual analysis, and avoid restricting inference to the groups sampled in the data and the

inclusion of a large number of dummy variables as in a fixed effects model. Multilevel models also offer more flexibility than other methods to correctly account for the complex structure of the social world.

The general form of a logistic multilevel model for purely hierarchical data with two levels is:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X}_{ij} + u_j. \qquad (1)$$

Here π_{ij} is the probability of individual *i* in cluster *j* taking on a value of 1 for the *y* variable, where *y* is a dummy variable indicating whether a person experienced an event or has a particular characteristic. β_0 represents the overall intercept in the linear relationship between the log-odds of *y* and the predictor variables included in the model, $\boldsymbol{X}_{ij}$, and is the log-odds for an individual pertaining to the reference categories of the categorical variables, having a value of 0 on continuous variables and belonging to the average higher-level group (a group with a value of 0 for the higher-level random effect u_j). The vector $\boldsymbol{\beta}_1$ contains the coefficients for each explanatory variable in the model when all other predictor variables are controlled for. These coefficients are also known as the cluster-specific effects of the explanatory variables, since they represent the effect of a unit increase in the covariate on the log-odds that the individual has a value of 1 on the outcome variable *y*, for a constant value of u_j and therefore within the same higher-level group *j*. The vector $\boldsymbol{X}_{ij}$ represents the predictor variables which may be defined at the individual or cluster level. The predictor variables may also include interaction effects or cross-level interaction effects. The u_j represent the random effects for the higher level classification units, which are assumed to follow a normal distribution with mean 0 and variances σ_u^2.

Besides purely hierarchical structures, multilevel models can also deal with data pertaining to two different non-hierarchical classifications (cross-classifications) (Fielding & Goldstein, 2006). The general form of such a cross-classified multilevel logistic model is:

$$\log\left(\frac{\pi_{i(js)}}{1-\pi_{i(js)}}\right) = \beta_0 + \boldsymbol{\beta}_1^{T}\boldsymbol{X}_{i(js)} + u_j + v_s. \qquad (2)$$

Here $\pi_{i(js)}$is the probability of individual *i* in clusters *j* and *s* taking on a value of 1 for the *y* variableThe parameters u_j and v_s represent the random effects for each higher-

level classification, which are assumed to follow a normal distribution with variances σ_u^2 and σ_v^2.

2.2. Review of Properties of Cross-classified Models

For the case of cross-classified multilevel models, sample-size requirements and the level of interpenetration required between the two cross-classified higher level classifications necessary for accurate parameter estimation have not yet been considered. What is currently available is a software package which produces power calculations for various sample sizes, data structures and random effects sizes – MLPowSim (Browne and Golalizadeh, 2009). For cross-classified models the estimation is carried out in R using the *lmer* function. The most flexible template for cross-classified data in MLPowSim enables the user to specify the total sample size, the number of higher-level groups, the probabilities of sampled cases pertaining to each higher-level combination, and the expected variances. The MLPowSim manual includes an example with exam attainment at age sixteen – a continuous variable – chosen as the outcome variable, where each student is associated with both a primary and secondary school. For this particular application, results show that sampling additional cases (students) from new higher-level groups (schools) results in greater power increases than sampling additional cases from higher-level groups already included in the sample. Also, adding further cases per higher-level grouping only benefits power calculations up to a threshold number of cases.

A number of papers exist (Rodriguez & Goldman, 1995; Paccagnella, 2011; Moineddin et al., 2007; Theall et al., 2011; Maas & Hox, 2005) that assess the impact of various factors, including sample size and outcome probability, on the properties of two-level hierarchical model estimates – for both continuous and binary outcome variables – through simulation studies. By definition two-level models do not include data pertaining to two classifications, and therefore the impact of interpenetration on model estimates cannot be reviewed in these studies. The results of these studies will be presented in the discussion section when reviewing this paper's results for cross-classified models.

2.3. Estimating Area and Interviewer Effects

The analysis of interviewer effects has become a popular application of multilevel methods (Von Sanden, 2004). Sample cases are nested within interviewers. However,

interviewers generally work in a limited geographic area, and to the extent that people from certain areas are more or less likely to cooperate, significant interviewer effects may simply indicate area effects. Moreover, there may also be area effects on nonresponse, arising due to similarities in socio-economic and cultural characteristics, in the perception of privacy, crime and safety, as well as in environmental factors such as physical accessibility and urbanicity across geographic boundaries (Haunberger, 2010). One approach to estimate interviewer effects in the past has been to simply ignore area effects (Pickery & Loosveldt, 2002; Haunberger, 2010; Blom et al., 2010) which clearly could yield misleading results and may overstate the effect of interviewers. Few studies have attempted to disentangle interviewer and area effects by specifying a cross-classified multilevel model for multistage cluster sample design data (Campanelli & O'Muircheartaigh, 1999; Durrant et al., 2010). This model specification prevents confounding only when the data is partially interpenetrated, which means that interviewers are not fully nested within areas, with interviewers working in more than one primary sampling unit (PSU), and cases in one PSU designated to more than one interviewer. In this particular application of sample units within interviewers operating within sampling areas, the higher-level variance is divided into two parts – the interviewer-level variance and the area-level variance.

3. Design of the Simulation Study

First in this section the process by which the data in the simulation study is generated is presented. Then, the cross-classified multilevel logistic regression model which is fitted to the simulated data is presented. The various simulation scenarios and the design factor values considered are then specified. Finally, the measures used to assess the properties of the estimators and test statistics, including the rationale for considering each measure and the equations used for their calculation, are presented.

3.1. Data Generating Procedure

In this simulation study the focus is on the random parameter estimates, and therefore only an overall intercept β_0 is included as a fixed effect. Its value is determined after considering the overall probability of the outcome for the mean area and the mean interviewer, π, by using the following formula:

$$\beta_0 = \log\frac{\pi}{1-\pi}. \qquad (3)$$

This value is fixed for all cases. Then a cluster-specific random effect for each interviewer and area is generated separately from a normal distribution of mean 0 and variances σ_u^2 and σ_v^2 respectively. The log-odds of each case, $\eta_{i(js)}$, are computed by adding the overall intercept value to the simulated random effects. These values are then converted to probabilities using the equation:

$$p_{i(js)} = \frac{\exp(\eta_{i(js)})}{1 + \exp(\eta_{i(js)})}. \qquad (4)$$

Values of the dependent variable $Y_{i(js)}$, a dichotomous outcome – with 0 signifying nonresponse and 1 signifying response to the survey request – for each case, are generated from a Bernoulli distribution with probability $p_{i(js)}$.

For scenarios which vary only in the interviewer case allocation the same set of 1000 cluster-specific random effects is used. This strategy underlies the fact that while interviewers are assumed to come from an infinite population, the allocation of workload from different areas to specific interviewers is limited to a finite number of possibilities.

3.2. Estimation of the Multilevel Cross-classified Model

The following multilevel cross-classified model is then fitted to the simulated data to identify interviewer and area random effects (without covariates for simplicity):

$$\text{logit}\left(p_{i(js)}\right) = \eta_{i(js)} = \beta_0 + u_j + v_s, \qquad (5)$$

where the interviewer-specific residuals u_j are distributed N(0, σ_u^2) and the area-specific residuals v_s are distributed N(0, σ_v^2). The analyses of the simulated datasets are carried out using STATA Version 12 calling MLwiN Version 2.25 through the 'runmlwin' command (Leckie & Charlton, 2011). Models are fitted using the Markov Chain Monte Carlo (MCMC) estimation method with default priors, a burn-in length of 10,000 and 200,000 iterations. Initial values for parameters are obtained by making use of the second order penalised quasi-likelihood (PQL) estimation method.

3.3. Simulation Scenarios

To explore the properties of estimators, a simulation experiment is carried out using a factorial design. The simulated scenarios vary in the following factors: overall sample size (N), number of interviewers and areas (N^I and N^A), and by consequence number of cases per interviewer and per area, level of cross-classification between interviewer

and area allocations, higher-level variance, and overall probability of the outcome variable (π).

The choice of the values for the various factors reflects realistic representations of general household survey scenarios. N^A in this simulation study will not be altered for a specific N. The initial numbers chosen for N, N^A, and N^I are based on the values obtained from a real survey and slightly adapted to obtain numbers which are easily divisible to obtain balanced designs. The main design, which will be referred to as the baseline scenario design, includes 120 areas consisting of 48 cases per area allocated to 240 interviewers who each have a workload of 24, totalling 5760 cases, with the area variance σ_v^2=0.3, interviewer variance σ_u^2=0.3 and an overall probability π =0.8. The impact of different interviewer-area classifications – varying in terms of the number of areas each interviewer works in (and consequently the number of interviewers per area) and the overlap of interviewers working in neighbouring areas – on the properties of the estimators and test statistic for the baseline scenario factors is analysed. The number of areas each interviewer works in will be allowed to vary from 1 to 6.

For illustration, the diagrams show the area-interviewer allocations for the first 6 areas. The areas are considered as sequential numbers in a circle, with the final area – area 120 – neighbouring the first area – area 1. Each box represents an area and the numbers within each box represent the interviewers working within that area (numbers from 1 to 240). The simplest case – CASE 1 – is where two interviewers work in each area, with each interviewer working only in one area (Diagram 1). In this case, there is no overlap in neighbouring areas with respect to the interviewers working within them. This in fact represents a purely hierarchical model, with individuals nested in interviewers which in turn are nested in areas.

Next, an interviewer can work in two areas, with four interviewers working in each area (Diagram 2). Three possible scenarios exist. The most overlap occurs for the scenario which allocates the same set of four interviewers to work in two neighbouring areas (CASE 2A). Alternatively, groups of three interviewers are repeated in two neighbouring areas with a fourth interviewer varying in the two areas (CASE 2B). Thirdly, pairs of interviewers are always allocated together, with each particular pair never occurring twice with another pair (CASE 2C).

Similar allocation patterns are considered for schemes where the interviewer works in three or more areas (diagrams not presented). The same basic principle of decreased overlap as one moves from the allocation A to subsequent allocations applies. For cases where interviewers work in three areas and each area includes six different interviewers, seven different allocation possibilities are considered. With interviewers working in more areas, less variations of overlap are considered, and this is simply due to the feasibility of such allocation schemes in practice. Three allocation schemes are considered for situations when each interviewer works in four, five and six areas, and cases within each area are allocated to eight, ten and twelve different interviewers respectively.

Due to computer power limitations and dependencies between factors – such that for a fixed sample size a change in the number of clusters (interviewers or areas) also changes the number of cases per cluster and the level of cross-classification between the two higher-level classifications, it was impossible to consider all factor combinations. Only one simulation factor at a time is changed, keeping all other factors constant. Table 1 outlines the baseline values as well as the other values considered for each factor in the simulation study.

The analysis for the initial baseline scenario design, containing 5760 cases, highlights a need to consider a smaller *N*. New datasets, amounting to one half and one fourth of the original baseline scenario caseload (2880 cases from 60 areas allocated to 120 interviewers and 1440 cases from 30 areas allocated to 60 interviewers) are also generated. For the baseline scenario there are twice as many interviewers as there are areas, $N^I=2N^A$. Another alternative considered is to have an equal number of interviewers and areas, $N^I=N^A$, that is, 120 interviewers for 120 areas for *N*=5760. For this data structure only six interviewer-area allocation schemes are considered, varying from the most geographically restrictive case where one interviewer works only in one area, to the most sparse where each interviewer works in six areas. In this case, variations in the amount of overlap in the groups of interviewers allocated to each area are not attempted, and the allocation schemes always allow the same group of interviewers to work together in neighbouring areas. These allocation schemes shown in Diagram 3, denoted as CASE *a*, where *a* represents the number of areas each interviewer works in, are therefore comparable to the allocation schemes CASE 2A outlined above.

3.4. Evaluation: Properties of the Estimators and Test Statistics

The models are assessed in terms of the following properties: the correlation of the two variance estimators, the percentage relative bias, the mean squared error, the standard error, the confidence interval coverage, and the power of tests. The covariance between the area and interviewer variance estimators is a quality measure in itself. For easier interpretation the correlation ρ for each dataset is calculated using the formula

$$\frac{1}{1000}\sum_{i=1}^{1000}\mathrm{corr}_i\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)=\frac{1}{1000}\sum_{\mathrm{i}=1}^{1000}\frac{\mathrm{cov}_{\mathrm{i}}\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)}{\sqrt{\mathrm{var}_i\left(\widehat{\sigma_u^2}\right)\mathrm{var}_i\left(\widehat{\sigma_v^2}\right)}}. \qquad (5)$$

'Good' estimators are expected to show no substantial correlation. High negative correlation values indicate problems with the identifiability of the two variance parameter estimates. In such cases the model may correctly estimate the total higher-level variance, which is the sum of the interviewer and area variances, but incorrectly apportion the variance to the two higher-level classifications, producing biased estimates for the individual random parameters. One estimate would be over-estimated, and the other estimate would be under-estimated, resulting in a negative correlation. Negative correlation values of -0.1 or higher will be considered problematic. Browne et al. (2001) make reference to this problem, and refer to it as the collinearity of random terms, and identify "poor mixing properties and high negative cross-chain correlations" (p.14) as good identifiers of this problem.

The percentage relative bias of the parameter is calculated to determine the accuracy of the parameter estimators. The model estimates are expected to always vary slightly from the true parameter value. Therefore, only percentage relative bias values above 3% will be considered substantial. Standard error accuracy is assessed using the coverage method (Maas & Hox, 2005), where coverage of the true parameter value within the 95% Wald confidence interval of the parameter estimate for each simulated dataset is recorded separately. The coverage rate is recorded for all simulation scenarios and compared with the nominal rate of 95%. The mean standard error for the parameter estimators gives an indication of the precision of the estimates for the various survey conditions. The null hypothesis, specifying the true parameter

value to be zero, is tested for both variance parameters of each simulated dataset by using the Wald test. The power of a test indicates the probability that the null hypothesis is correctly rejected. Maas and Hox (2005) explain that basing the testing of significance for variance parameters using the asymptotic standard error is not ideal. Such a test is based on normality assumptions. Testing of the null hypothesis, which specifies the random parameter to be equal to zero, lies on the boundary of the permissible parameter space, since variances can only be positive. The standard likelihood theory no longer holds at this boundary. However, this practice is widely used and justifies its use in this simulation study. In calculating the power for the variance parameters the p-values are halved, since variances cannot be negative, and therefore the alternative hypothesis is one-sided (Snijders & Bosker, 1999).

4. Results

To remind the reader of what has been outlined in the design section, the baseline scenario design has the following properties: 120 areas (48 cases per area) allocated to 240 interviewers (24 cases per interviewer), totalling 5760 cases, σ_v^2=0.3, σ_u^2=0.3 and π=0.8. Generally one or two factors from the following: σ_v^2 and σ_u^2, π, *N* and the ratio of interviewers to areas (dependent on N^I and N^A), are changed for every new scenario. For every specific set of factor values different interviewer allocation schemes are specified, giving rise to more scenarios. The impact (or lack of effect) of these factors on the properties of the estimator are reviewed. General patterns are documented and any possible interactions between factors highlighted.

The properties for the overall intercept β_0 showed relatively stable results across different factor values. Under all simulation scenarios the test for β_0 obtains a power of 1. Accurate intercept estimates $\hat{\beta}_0$ are obtained even for small *N* and very geographically-restrictive interviewer allocation schemes. The highest absolute relative percentage bias for the β_0 estimator is less than 0.6%. This slight deviation of the mean estimate from the true parameter may simply reflect small sample bias rather than any methodological bias. The Wald coverage rates are close to the 95% nominal rates across all scenarios. Consequently, the analysis of the impact of various factor changes for the above-mentioned properties will be restricted to the random parameters. On

the other hand, the standard error of the fixed effect estimator shows some variation across factors. These patterns for each property across factors will now be summarised.

4.1. Power of Test

For the baseline scenario design the power is higher than 0.9 for both random parameters for all case allocation schemes (Table 2, Columns 1 & 2). The power of the Wald test at the 5% significance level is close to the optimal value of 1 for all interviewer case allocation possibilities for both random parameters except for the test for σ_v^2 for the least sparse interviewer allocation (CASE 1) which yields a power of 0.91 (Table 2, Columns 1 & 2).

Interviewer dispersion is the factor which shows the greatest impact on the power. For scenarios with one interviewer per area allocation scheme power is observed to decrease to 0 for certain scenarios, whereas the lowest power observed for two interviewers per area allocation schemes is 0.67. There is a threshold, which varies for different factor value combinations, beyond which further dispersion does not yield power gains. On the other hand, reduced interviewer overlap for a constant number of areas per interviewer does not improve the power. Here overlap refers to the extent that the group of interviewers working in neighbouring areas are the same, such that CASE 2A has greater overlap than CASE 2C.

Table 2 shows that for scenarios with smaller *N*, but keeping constant all other factors, lower power is obtained for the allocation schemes with the least interviewer dispersion (number of areas an interviewer works in). Therefore, sparser interviewer allocation schemes are required to obtain similar high levels of power for scenarios with a smaller *N*. The effect of sample size reduction on power is greater for N^I=N^A scenarios (Table 2, Columns 1–6) compared with N^I=2N^A scenarios (Table 2, Columns 7–12).

When only the overall probability (π=0.7, 0.8, 0.9) varies, with other factors kept constant at their baseline values, for CASE 1 scenarios higher overall probabilities result in lower power for the random parameters σ_v^2 and σ_u^2. High overall probabilities seem to have a greater impact on the power of tests for random effects parameters

which have a smaller number of higher-level units in the sample, i.e. the area random parameter σ_v^2 compared to the interviewer random parameter σ_u^2.

The number of higher-level units as well as interviewer dispersion mediate the effect of a lower variance on the power of the test for the random parameter, such that the only difference in power across different variances is observed for the one area per interviewer allocation for the area variance parameter. For the scenario with smaller area variance (σ_v^2=0.2) the power of the test for the area random parameter for the most geographically restricted interviewer allocation (1 area per interviewer) is substantially lower at 0.68 than the power for the baseline scenario design of 0.91. Increasing the area variance σ_v^2 to 0.4 improves the power for the CASE 1 allocation scheme from 0.91 to 0.99. On the other hand, for the scenario with smaller interviewer variance (σ_u^2=0.2), but keeping constant all other factors, the power of the test for the interviewer random parameter for CASE 1 is 1.

For N^I=2N^A scenarios, where substantial differences can be noticed for the power of the tests for the random parameters, the power for the area parameter σ_v^2 is consistently lower than that for the interviewer parameter σ_u^2 (Table 2). No difference is observed for N^I=N^A scenarios. These results indicate that the number of higher-level units mediates the effect of a lower intra cluster correlation (ICC) on the power of the tests for the random parameters.

The ratio of interviewers to areas also influences the power for the random parameters. Scenarios having N^I = N^A require more interviewer dispersion than equivalent N^I=2N^A scenarios to obtain the same power for the random parameters (Table 2).

4.2. Correlation between Random Parameter Estimators

Interviewer dispersion highly influences ρ between the two variance estimators. High negative correlations (greater than 0.4 and up to a maximum of 0.91) are obtained for all scenarios when interviewers are working in only one area. This correlation is reduced to less than -0.2 once interviewers work in two areas (Table 3).

No effect of sample size on ρ is observed for allocation schemes which allocate interviewers to at least two areas (Table 3). For N^I=2N^A scenarios ρ varies only from -

0.45 and -0.46 for the 5760 and 1880 sample size scenarios to -0.40 for the 1440 sample size scenario for CASE 1. For N^I=N^A scenarios when varying total sample size but keeping other factors constant (σ_v^2=0.3, σ_u^2=0.3, π=0.8), ρ decreases from 0.91 to -0.83 to -0.69 for the 5760 cases, 2880 cases and 1440 cases for CASE 1 (Table 3, Columns 4-6, Row 1). For more sparse allocation schemes no substantial differences can be observed for different *N* scenarios. Therefore, the effect of *N* on ρ is mediated by the number of higher-level units, or an unequal ratio of the two higher-level units, as well as the interviewer dispersion.

Scenarios with equal numbers of areas and interviewers obtain higher negative correlations than scenarios with twice the number of interviewers to areas (Table 3). This difference may be explained in terms of improved identifiability of the variance decomposition for scenarios with higher number of clusters, or alternatively an unequal number of clusters for the two classifications.

The negative correlation increases with increasing overall probabilities (Table 4). The increase in ρ from scenarios with π=0.8 to those with π=0.9 is greater than the increase from scenarios with π=0.7 to those with π=0.8, indicating that the effect of π on ρ is monotonic but not linear. For allocation schemes with at least three areas per interviewer the effect of overall probability on ρ is no longer present.

Higher area variance values result in lower negative correlations for the more restrictive interviewer allocation schemes. No trend is identified when varying the interviewer variance. These results suggest that the number of higher-level units associated with a variance parameter mediates the effect of the variance on ρ. Lower negative correlation is obtained for the two areas per interviewer allocation schemes which have less overlap.

The effect of interviewer overlap is no longer present for more dispersed interviewer allocation schemes. This result indicates that the impact of interviewer overlap is mediated by the interviewer dispersion, that is, the number of areas an interviewer works in.

4.3. Percentage Relative Bias of Parameter Estimators

In most scenarios with *N*=5760, the relative percentage biases for the variance parameter estimators are around 1-3% once interviewers are allocated work in at least

two areas (Table 5, Column 1). The bias is much higher for interviewer allocation schemes which restrict the interviewer to working in one area (CASE 1). The biases for CASE2-6 fluctuate around within the range specified above, failing to show any systematic reduction with further dispersion and less interviewer overlap. For interviewer case allocation schemes in which interviewers are working in at least two areas, the area random parameter σ_v^2 bias is almost always greater than the interviewer random parameter σ_u^2 bias (Table 5, Columns 1-6, Rows 2-6). This again confirms the importance of group size for the accuracy of parameter estimators.

As expected, greater biases for the σ_v^2 and σ_u^2 estimators are observed for smaller *N*, with the scenario including 1440 cases with $N^I = N^A$ obtaining biases between 5-13% for all allocation schemes (Table 5, Column 12). Scenarios with N^I=N^A (Table 5, Columns 7-12) generally obtain higher biases for both variance parameter estimators than N^I=$2N^A$ scenarios (Table 5, Columns 1-6). This trend is observable for the interviewer parameter σ_u^2 estimator. This trend is what would be expected due to the greater number of interviewers in the N^I=N^A scenarios compared to the N^I=$2N^A$ scenarios. On the other hand, for the area parameter σ_v^2 estimator – where N^A=120 in both the N^I=N^A and N^I=$2N^A$ scenarios – this pattern is less consistent for the 5760 and 2880 sample size scenarios. However, with a total sample size of 1440 the N^I=N^A scenario yields consistently higher biases than the N^I=$2N^A$ scenarios. These results may support the post-hoc hypothesis that having an unequal number of clusters (interviewers and areas) also improves the quality of estimates, albeit not as strongly as increasing the number of groups in each higher-level classification.

No clear trend for the change in bias by interviewer overlap, interviewer dispersion beyond two areas per interviewer, overall probability and by variances is observed.

In this study the percentage relative bias of the MCMC posterior median has also been calculated. On the whole, the biases for the posterior mean and the posterior median show similar trends as the factor change. One particular difference is the lower bias obtained for the estimators based on the 50% percentile in comparison to the estimators based on the mean for scenarios with smaller *N* and N^I=N^A.

4.4. Wald Confidence Interval Coverage

The Wald confidence interval coverage rates are close to 95% nominal rate – between 94–96% – in most scenarios. However, there are some cases of under-coverage (lowest observed rate is 87%) as well as very few cases of over-coverage (highest observed rate is 100%) for scenarios where each interviewer works only in one area.

Slightly lower coverage rates are observed for smaller *N* in most scenarios for both σ_v^2 and σ_u^2. Only the scenarios with the smallest sample size of *N*=1440 consistently obtain non-accurate coverage rates across all interviewer case allocation schemes. However, these rates do not fall below 89% once each interviewer is allocated work in at least two areas. Coverage rates closer to the 95% nominal rate for the σ_u^2 parameter are noticeable for the N^I=2N^A scenarios compared to the N^I=N^A scenarios for *N*=5760. This improvement in the confidence interval coverage rate with an increase in the number of interviewers from 120 interviewers to 240 interviewers no longer occurs for smaller *N*.

Some factors considered in this study do not seem to influence coverage rates. There does not seem to be a consistent pattern in the coverage rates by the overall probability or by the higher-level variances. Neither do the results show any evidence of the extent of interviewer overlap influencing coverage rates. Unexpectedly, the results do not provide any evidence that the MCMC credible quantiles perform consistently better than the intervals based on asymptotic normality. This result may reflect the fact that the values for the variances considered in the simulations are not close enough to zero. Had smaller variances been considered, possibly an improvement in the coverage of the MCMC credible quantiles in comparison to the Wald confidence interval may have been observed.

4.5. Standard Errors

The precision of both fixed effect and random effects estimators is affected by *N* (Table 6). As expected, reducing the sample size to one fourth of the original *N* (from 5760 cases to 1440 cases) seems to approximately double the standard errors for all estimators. For the N^I=2N^A scenarios the σ_v^2 estimator obtains higher standard errors than the interviewer variance estimator, thus highlighting the positive impact of a

higher number of clusters on the precision of the estimator. As expected, there is no substantial difference in the standard error of the two variance estimators for the N^I=N^A scenarios.

The standard errors of the variance estimators decrease with greater interviewer dispersion, up to a threshold number of areas per interviewer, which varies by *N* and the ratio of interviewers to areas (Table 6). Higher standard errors are obtained for scenarios with N^I=N^A compared to scenarios with N^I=2N^A. This result highlights the increased precision for scenarios with unequal number of higher-level units for the two higher-level classifications. Table 6 shows that the discrepancy in the standard errors for N^I=N^A scenarios compared to the N^I=2N^A scenarios are more pronounced for more geographically restricted interviewer allocations, indicating that to some extent interviewer dispersion mediates the effect of the number of higher-level units on the standard error of the estimator.

Interviewer overlap does not seem to affect the size of the standard errors. A higher overall probability results in higher standard errors for all three parameter estimators, with some increase from π=0.7 to π=0.8, and a much higher increase from π=0.8 to π=0.9, especially for the CASE 1 interviewer case allocation scheme. This non-linear result is similar to the effect of overall probability on ρ, which shows that a greater increase in ρ between the two estimators is observed for the extreme end of the probability scale, when increasing the overall probability from 0.8 to 0.9. When the value of the variance changes, the standard error changes in the same direction for the respective variance estimator. The unchanged variance does not experience changes in its estimator's standard errors, once interviewers work in at least two areas.

5. Discussion of Results

In this section the findings above are discussed in their wider context, and comparisons to findings in other studies are made. Initial implications for survey designs are highlighted. As expected, the results show worse quality estimators for smaller *N*. It is important to consider that in this study it is not possible to clearly distinguish between the effects of decreases in *N* and decreases in N^A and N^I, since halving the *N* also reduces the number of higher-level units by half. Consequently, the

results of halving the *N* while keeping the same N^A and N^I (by reducing the cluster sizes) have not been assessed. Bias has been found to increase with decreases in *N*, and this increase is consistent for all interviewer case allocation schemes considered in the study. The greatest increase in bias with smaller *N* is observed for CASE 1. Allocating each interviewer cases in two different areas reduces the effect of smaller *N* on bias. However, sparser allocation schemes do not seem to mediate this effect further. The increases in the biases are particularly pronounced when halving *N* from 2880 to 1440 for the $N^I = N^A$ scenarios. This is similar to the result obtained by Paccagnella (2011) who shows that the improvements in the estimators' accuracy with sample expansions decrease as *N* increases. Similarly to Moineddin et al. (2007), there is some evidence in this study of lower coverage rates for smaller *N*. The confidence interval coverage rates are slightly lower for the 1440 sample size scenario compared to the 5760 and 2880 sample size scenarios for all interviewer case allocation schemes. Power also decreases for smaller total sample sizes. However, for the 2880 sample size scenarios this decrease can only be noticed up to two areas per interviewer allocation schemes for the $N^I=2N^A$ scenarios and three areas per interviewer allocation schemes for the $N^I=N^A$ scenario. For the 1440 sample size scenarios the power values are lower compared to the 2880 sample size scenario for all interviewer case allocation schemes, and even for 6 areas per interviewer allocation schemes power ranges from 0.89 to 0.92. The opposite trend can be observed for the correlation between the two random parameter estimators, with the one area per interviewer allocation scheme showing a decrease in the negative correlation with decreasing *N*. This trend is more pronounced in the $N^I=N^A$ scenario than the $N^I=2N^A$. However, this trend is negligible for both these scenarios once interviewers are working in at least two areas each. Standard errors of both the overall intercept and random parameter estimators seem to increase monotonically with decreasing *N*. Interviewer dispersion does not mediate the effect of decreasing *N* on standard errors. However, for a constant *N* the precision of variance estimators improves with further interviewer dispersion – up to a limit of 3 areas per interviewer – for $N^I=2N^A$ scenarios.

The above-mentioned results on the relationship between *N* and the various properties show that reductions in *N* can be mediated to some extent by interviewer dispersion. However, small *N* – 1440 cases – are to be avoided as even with sparse interviewer allocation schemes they do not achieve acceptable levels of accuracy,

precision and power. On the other hand, large and medium sized samples, including N^I=2N^A scenarios, obtain good estimates once interviewers work in at least three areas. The percentage relative bias does not fall below 1%, even for the largest sample considered (5760 cases). Estimators of higher-level parameters obtain bias values of up to 3% even for large *N* and a large number of higher-level units (240 interviewers, 120 areas). This is similar to the results presented by Moineddin et al. (2007), where for data with 100 groups of size 50, bias levels for random effects estimates are all under 4%, but never reach 1% or lower.

The comparison of the N^I=2N^A with the N^I=N^A scenarios indicates that a higher number of clusters as opposed to a higher cluster size for a constant *N* yields better estimates. In this paper, the *N* does not increase as the number of groups is increased. Instead, the number of groups is altered for a set *N*. Lower negative correlation between the two higher-level random effects, higher power for the Wald test for σ_u^2, lower standards errors for $\widehat{\sigma_u^2}$ and lower relative percentage bias for $\widehat{\sigma_u^2}$ are observed for the N^I =2 N^A compared with the N^I = N^A scenarios for some of the least sparse interviewer allocation schemes, and especially for smaller *N*. The improvement in the accuracy and precision of $\widehat{\sigma_v^2}$ for the smallest sample size scenario and the higher power for the Wald test for σ_v^2 may be indicating that besides the effect of the number of clusters (which for the area classification remains unchanged), the ratio of higher classification units may also affect the quality of estimates with a ratio unequal to one performing better. This result suggests that a larger N^I should be working within a set N^A for best quality estimates.

These results are consistent with previous simulation studies for two-level hierarchical models which emphasise the importance of a higher number of clusters, as opposed to a larger cluster size, for the quality of estimates from multilevel models. Maas and Hox (2005) find that the coverage rates for variance parameters only increase with increases in the number of groups, and show no change for increasing group size. Paccagnella (2011) only documents a decrease in bias for the variance components estimators with an increase in the number of groups, despite the fact that both the group size and the number of groups are included as varying factors in their simulation study. Mok (1995) looks specifically at comparing the bias for estimators from 2-level models when simulating data with different designs, comprising different

student (level 1) to school (level 2) ratios for various fixed *N*. Type a designs have a ratio of students per school over number of schools greater than 1; Type b designs have an equal ratio; and Type c designs have a ratio of less than 1. Mok (1995) concludes that for a fixed *N*, larger standard errors and larger mean squared errors are obtained for Type a designs compared to Type b and c designs for the variance estimator, but she finds no association between design type and bias for the random intercept estimator. Moineddin et al. (2007) find that both the group size and the number of groups affect the accuracy of random parameter estimates. Very small group sizes of 5 give very high biases. However, for a scenario including 30 groups of size 30 each, an increase to 50 groups leads to a larger decrease in bias compared to an increase to a group size of 50. On the other hand, the number of groups is positively related to the confidence interval coverage rates for both the random intercept and the random slope parameters, whereas the group size is only significantly related to the coverage rates for the random slope parameter. Rodriguez and Goldman (1995) find both higher bias and inflated standard errors for variances of higher-level classifications with small cluster sizes. In this study the implications of small group sizes have not been explored since sampling very small numbers from a sampling area is not common practice due to survey travelling costs and other administrative expenses. While it is possible to envisage a few interviewers having a very small caseload in very remote areas, the majority of interviewers are generally assigned a bigger caseload.

In this study lower power of the Wald test for the random parameters and higher correlation between the two random parameter estimators are found for higher overall probabilities for some restrictive interviewer case allocation schemes. Higher standard errors are obtained consistently for all estimators across all interviewer case allocation schemes for higher overall probabilities. Moineddin et al. (2007) find that for 2-level models lower prevalence rates of 0.1 result in higher bias and lower coverage rates compared to higher overall probabilities. Moineddin et al. (2007) use the values 0.1, 0.34 and 0.45 for the overall probabilities. In this study the values 0.7, 0.8 and 0.9 are included in the analysis. Both studies suggest that estimates of lower quality are obtained for extreme values, with Moineddin et al. (2007) investigating the lower end of the spectrum and this study investigating the higher end. For the scenarios considered the negative correlation between the two random parameter estimators is

reduced to less than 0.1 once the interviewers were allocated work in three areas. Moreover, the effect of the overall probability on this correlation is only observed up to interviewer allocation 3A. In the case of the effect of the overall probability on the power of the Wald test, this is restricted to just the most restrictive interviewer case allocation – CASE 1. Once interviewers work in two areas, no effect of the overall probability on power is observed. Consequently, some of the effects of the overall probability on the quality of estimates can be avoided during the survey administration by assigning work to interviewers in at least three areas.

There are mixed results in the literature on the effect of ICC on the quality of parameter estimates. Random intercept estimators have been shown to differ significantly by ICC values in Moineddin et al. (2007), showing higher bias for lower ICC values. Moineddin et al. (2007) also observe a trend of higher coverage rates for higher ICC values for the random intercept. On the other hand, Maas and Hox (2005) and Paccagnella (2011) do not find a significant effect of the ICC value on the relative bias or the Wald 95% confidence interval coverage rates for random parameters.

Similarly, in this study the size effect and direction of the effect of ICC on the quality of the estimates seems to vary for different properties. Higher ICC values seem to decrease the negative ρ, although this is no longer present for higher-level effects with a large number of clusters in the sample. In fact, lower negative ρ are observed for higher area variances σ^2_v up until interviewer allocation CASE 3A, but no consistent change is observed for higher interviewer variances σ^2_u in scenarios with double the number of interviewers to areas. Similarly, the ICC is found to have a positive relationship with the power of the Wald test for the most restrictive interviewer case allocation, CASE 1, but again for the other higher-level classification with twice the number of clusters this effect is not observed. In contrast, precision seems to decrease for higher variances. Similarly to Maas and Hox (2005) and Paccagnella (2011), in this study no clear pattern for the change in the percentage relative mean bias of the variance parameter estimators by ICC is observed. Contrary to the results reported by Moineddin et al. (2007), in this study no evidence of the effect of ICC on the confidence interval coverage rates has been found. Similar to the effect of overall probability on the quality of estimates, these results indicate that generally once each interviewer is allocated cases in two, and sometimes, three different areas, small ICC

values will not be detrimental to the quality of the estimates. It is important to consider that in this study very small variances are not being investigated.

Interviewer dispersion, which refers to the number of areas each interviewer works in, only improves the quality of estimates up to a point. The power of the Wald test at the 5% significance level for the baseline scenario design is close to the optimal value of 1 for all interviewer case allocation schemes. For scenarios with smaller *N*, but keeping constant all other factors, sparser interviewer allocation schemes are required to obtain high power. Improvements in power are observed when increasing the number of areas per interviewer from one to two for *N*=2880 and *N*=1440, and from two to three for *N*=1440. Further dispersion only yields very small gains. The correlation between the two parameter estimators is reduced to the chosen threshold of -0.1 once interviewers are allocated to two areas for N^I=2N^A scenarios, and three areas for $N^I = N^A$ scenarios. More sparse allocation schemes do not result in substantially lower ρ for the scenarios considered. Decreases in the relative percentage bias are substantial when comparing the CASE 2 to the CASE 1 allocation scheme. However, no systematic trend in bias reduction is observed for CASE 3-CASE 6. Confidence interval coverage rates show problems of over- and under-coverage for different scenarios with the CASE 1 allocation scheme, but are close to the 95% nominal rate for all other allocation schemes. Standard errors for the variance estimators decrease with greater interviewer dispersion up to a certain number of areas per interviewer, which varies by *N* and ratio of interviewers to areas. For N^I=2N^A scenarios substantial decreases in standard errors are only present up to CASE 2 for N=5760, and CASE 3 for smaller *N*. $N^I = N^A$ scenarios show decreases in standard errors up to CASE 4 for *N*=5760 and *N*=2880, and CASE 5 for *N*=1440.

No consistent relationship between bias, confidence interval coverage rates, standard errors and power of the Wald test with the extent of interviewer overlap is found. The only impact of interviewer overlap was restricted to the ρ values for 2 areas per interviewer allocation schemes, with less overlap resulting in lower negative ρ. Consequently for the scenarios considered in this study, once all interviewers work in at least three areas, there is no benefit in aiming for less interviewer overlap. This result indicates that complicating interviewer case assignments by sending interviewers farther away from their area of residence in an attempt to avoid having the same

interviewers working in the same neighbouring areas is not necessary to obtain quality estimates.

6. Conclusions and Implications for Survey Design

The simulations in this paper offer new insight into the performance of the advanced multilevel models for realistic survey design conditions. This paper is the first work investigating the properties for cross-classified models under different survey design conditions. This paper has identified trends in the properties of the estimators and test statistics across a range of values for the simulation factors considered.

This paper indicates that, as expected, purely hierarchical data is subject to substantial biases, larger standard errors, high negative correlations between the two random parameter estimates, under and over coverage of the Wald confidence interval, and low power of the Wald test. Interpenetration of the higher-level groups at the design stage is required to allow for the two higher-level effects to be disentangled when estimating these effects using multilevel cross-classified models. For the scenarios considered in this paper limited overlap of the higher-level groups (of around 3 areas per interviewer for medium or large sample sizes) has been shown to provide sufficient interpenetration for good properties. Further dispersion yields only very small or negligible improvements in the properties. Overlap of the higher-level groups also acts as a mediating factor on the effect of the other simulation factors (sample size, the ratio of interviewers to areas, the overall probability, and the variance values) on the properties of the estimators and test statistic. Reductions in total sample size can be mediated to some extent by interviewer dispersion. However, small N – 1440 cases – are to be avoided as even with sparse interviewer allocation schemes they do not achieve acceptable levels of accuracy, precision and power. Importantly, the results also show that once interviewers work in at least three areas complicating interviewer case assignments by sending interviewers farther away from their area of residence in an attempt to avoid having the same interviewers working in the same neighbouring areas does not improve the quality of estimates. Consequently, study designs should focus on allowing some interpenetration between the two higher-level groups, whilst avoiding increasing survey costs or complicating logistics by disregarding the extent of overlap of higher-level units from the same classification

group. The results also suggest that for a fixed total sample size, a higher number of clusters as opposed to a higher cluster size yields better estimates. Moreover, the ratio of higher classification units may also affect the quality of estimates with a ratio unequal to one performing better.

It is acknowledged that the results from this paper are restricted to the factor values chosen and the scenarios considered. The results cannot necessarily be extrapolated to very different survey design conditions with certainty. Further research investigating different simulation factor values and data structures should be carried out to corroborate and extend existing evidence on the performance of these models. One particular area of further research should focus on the examination of these properties for very small higher-level variances. Although the factor conditions and the application considered here are specific to survey design and the exploration of interviewer effects on nonresponse, the same problem of identifiability may arise in other settings, such as in the investigation of the influence of community physiotherapists and the influence of the hospitals in the rehabilitation of patients having undergone orthopaedic surgery.

The paper considers the properties of variance estimators only. The data is generated from models including an overall intercept and random effects. No explanatory variables are considered. Other simulation papers reviewed earlier indicate that the worst estimator and test statistic properties are observed for the variance estimators. Consequently, the focus on the random effects is justified, as these parameters are the ones most susceptible to influence by changes in simulation factors. Moreover, scenarios achieving acceptable properties for the variance parameters can be assumed to also provide acceptable properties for fixed effect parameters. In future work the inclusion of fixed effects, especially cross-level interaction effects and contextual effects, should be considered.

This work created the procedure and R and STATA code that can be used independently of this research project to investigate the performance of multilevel cross-classified logistic models for existing data structures, or to inform the design of future studies with similar designs. A future project may focus on creating an online platform, similar to the MLPowSim tool (Browne & Golalizadeh, 2009), for other users

to be able to specify their data structure and run the simulation for their own specific application.

Acknowledgements

Vassallo's work was supported by the University of Southampton, School of Social Sciences Teaching Studentship and by the UK Economic and Social Research Council (ESRC), PhD Studentship (ES/1026258/1). Durrant's and Smith's work was supported by ESRC grant number RES-062-23: 'The Use of Paradata in Cross-Sectional and Longitudinal Research'.

References

Blom, A.G., De Leeuw, E.D. & Hox, J.J. (2010). Interviewer effects on nonresponse in the European Social Survey. *ISER Working paper Series, 2010-25, Institute for Social & Economic Research, ESRC.*

Browne, W. & Golalizadeh, M. (2009). *MLPowSim.* Centre for Multilevel Modelling, University of Bristol.

Browne, W. J., Goldstein, H. & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling, 1,* 103-124.

Campanelli, P. & O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey nonresponse. *Quality & Quantity, 33,* 59-76.

Durrant, G. & 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(2),* 361-381.

Durrant, G. B., Groves, R. M., Staetsky, L. & Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly, 74,* 1-36.

Fielding, A. & Goldstein, H. (2006). *Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review*. Research Report RR791. London, Department for Education and Skills. Retrieved May 18, 2011 from https://www.education.gov.uk/publications/standard/publicationDetail/Page1/RR791

Goldstein, H. (2011). *Multilevel Statistical Models.* Fourth Edition. Wiley, Chichester.

Haunberger, S. (2010). The effects of interviewer, respondent and area characteristics on cooperation in panel surveys: a multilevel approach. *Quality & Quantity, 44,* 957-969.

Leckie, G. & Charlton, C. (2011). *runmlwin: Stata module for fitting multilevel models in the MLwiN software package*. Centre for Multilevel Modelling, University of Bristol.

Maas, C. J.M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology: European Journal of Research Methods for the Behavioural and Social Science, 1,* 85-91.

Moineddin, R., Matheson, F. I. & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology,* 7, 34.

Mok, M. (1995). Sample Size Requirements for 2-level Designs in Educational Research. *Multilevel Modelling Newsletter, 7(2),* 11-15.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology, 7(3),* 111-120.

Pickery, J., Loosveldt, G. & Carton, A. (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys – A multilevel approach. *Sociological Methods & Research, 29,* 509-523.

Pickery, J. & Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. *Quality & Quantity, 36,* 427-437.

Rodriguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A,* 158, 73-90.

Schnell, R. & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics, 21(3),* 389-410.

Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: an introduction to basic and advanced multilevel modelling*. London: Sage.

Theall, K.P., Scribner, R., Broyles, S., Yu, Q., Chotalia, J., Simonsen, N., Schonlau, M. & Carlin, B. P. (2011). Impact of small group size on neighbourhood influences in multilevel models. *J Epidemiol Community Health, 65,* 688-695.

Von Sanden, N. D. (2004). *Interviewer effects in household surveys: estimation and design.* Unpublished PhD thesis, School of Mathematics and Applied Statistics, University of Wollongong. Retrieved February 24, 2012 from http://ro.uow.edu.au/theses/312/.

Tables and Diagrams

Table 1: Factor Values for Baseline and Other Scenarios

Factor	Baseline	Other
Number of cases per interviewer	24	48
Number of interviewers	240	30, 60, 120
Overall sample size	5760	1440, 2880
Overall propensity to respond	0.8	0.7, 0.9
Area variance	0.3	0.2, 0.4
Interviewer variance	0.3	0.2, 0.4

Table 2: Power of Wald Test at the 95% Confidence Level by Sample Size, Ratio of Interviewers to Areas and Interviewer Allocation (IA)

	$N^I=2N^A$	$N^I=N^A$
	Sample Size	

	5760		2880		1440		5760		2880		1440	
IA	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2
1	0.91	1.00	0.63	0.92	0.30	0.58	0.07	0.08	0.01	0.01	0.00	0.00
2	1.00	1.00	0.96	0.98	0.77	0.81	1.00	1.00	0.97	0.98	0.67	0.68
3	1.00	1.00	1.00	1.00	0.91	0.89	1.00	1.00	0.99	0.96	0.73	0.64
4	1.00	1.00	1.00	1.00	0.88	0.86	1.00	1.00	1.00	1.00	0.85	0.85
5	1.00	1.00	1.00	1.00	0.91	0.89	1.00	1.00	1.00	1.00	0.88	0.88
6	1.00	1.00	1.00	1.00	0.92	0.88	1.00	1.00	1.00	1.00	0.91	0.88

Constant factor values: σ_v^2=0.3, σ_u^2=0.3, π=0.8
$N^I=2N^A$: N^I=240 and N^A=120 for *N*=5760; N^I=120 and N^A=60 for *N*=2880, N^I=60 and N^A=30 for *N*=1440; $N^I=N^A$: N^I=120 and N^A=120 for *N*=5760; N^I=60 and N^A=60 for *N*=2880, N^I=30 and N^A=30 for *N*=1440

Table 3: ρ by Sample Size, Ratio of Interviewers to Areas and Interviewer Allocation

	$N^I=2N^A$			$N^I=N^A$		
	Sample Size					
IA	5760	2880	1440	5760	2880	1440
1	-0.45	-0.46	-0.40	-0.91	-0.83	-0.69
2	-0.09	-0.11	-0.09	-0.19	-0.17	-0.15
3	-0.03	-0.02	0.04	-0.13	-0.12	-0.11
4	0.01	0.01	0.00	-0.04	-0.04	-0.03
5	0.02	0.02	0.03	-0.02	-0.01	-0.01
6	0.03	0.03	0.03	0.00	0.00	0.01

Constant factor values: σ_v^2=0.3, σ_u^2=0.3, π=0.8, $N^I=2N^A$
N^I=240 and N^A=120 for *N*=5760; N^I=120 and N^A=60 for *N*=2880, N^I=60 and N^A=30 for *N*=1440

Table 4: ρ by Overall Probability and Interviewer Allocation

IA	Overall Probability

	0.7	0.8	0.9
1	−0.43	−0.45	−0.50
2A	−0.08	−0.09	−0.12
2C	−0.04	−0.05	−0.10
3A	−0.01	−0.03	−0.04

Constant factor values: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, $N^I=2N^A$

Table 5: Relative Percentage Bias by Sample Size, Ratio of Interviewers to Areas and Interviewer Allocation

	$N^I=2N^A$						$N^I=N^A$					
	$\widehat{\sigma_v^2}$			$\widehat{\sigma_u^2}$			$\widehat{\sigma_v^2}$			$\widehat{\sigma_u^2}$		
	Sample Size											
IA	5760	2880	1440	5760	2880	1440	5760	2880	1440	5760	2880	1440
1	−3.2	−6.7	−5.3	6.8	11.2	19.8	2.3	4.4	12.5	3.6	5.6	11.3
2	2.0	2.6	4.8	1.3	1.9	2.4	3.6	4.0	10.8	1.5	5.0	9.0
3	2.4	4.2	6.1	0.1	1.2	1.1	1.6	3.1	10.5	1.0	4.3	5.3
4	1.7	3.3	5.0	0.7	1.3	1.8	1.7	1.5	9.8	1.9	4.2	9.7
5	1.7	2.4	7.2	1.0	1.5	3.4	2.0	2.6	8.6	1.4	4.9	8.3
6	1.1	3.1	7.4	0.7	1.8	2.4	1.6	3.8	10.3	1.9	3.0	6.7

Constant factor values: σ_v^2=0.3, σ_u^2=0.3, π=0.8
$N^I=2N^A$: N^I=240 and N^A=120 for N=5760; N^I=120 and N^A=60 for N=2880, N^I=60 and N^A=30 for N=1440; $N^I=N^A$: N^I=120 and N^A=120 for N=5760; N^I=60 and N^A=60 for N=2880, N^I=30 and N^A=30 for N=1440

Table 6: Standard Errors by Sample Size, Interviewer Allocation and Ratio of Interviewers to Areas

	$N^I=2N^A$ scenarios					
	Sample Size					
	5760		2880		1440	
	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$
1	0.094	0.085	0.148	0.143	0.191	0.184
2	0.070	0.063	0.104	0.094	0.153	0.134
3	0.067	0.060	0.097	0.087	0.140	0.123
4	0.065	0.059	0.095	0.085	0.143	0.126
5	0.064	0.058	0.093	0.084	0.143	0.125
6	0.064	0.059	0.092	0.084	0.142	0.123
	$N^I=2N^A$ scenarios					
1	0.252	0.252	0.273	0.273	0.318	0.317
2	0.077	0.076	0.111	0.112	0.171	0.169
3	0.073	0.075	0.107	0.112	0.165	0.167
4	0.067	0.067	0.096	0.098	0.153	0.153

5	0.065	0.065	0.095	0.096	0.147	0.147
6	0.064	0.064	0.095	0.094	0.147	0.144

Constant factor values: σ_v^2=0.3, σ_u^2=0.3, π=0.8
N^I=2N^A: N^I=240 and N^A=120 for *N*=5760; N^I=120 and N^A=60 for *N*=2880, N^I=60 and N^A=30 for *N*=1440; N^I=N^A: N^I=120 and N^A=120 for *N*=5760; N^I=60 and N^A=60 for *N*=2880, N^I=30 and N^A=30 for *N*=1440

Diagram 1

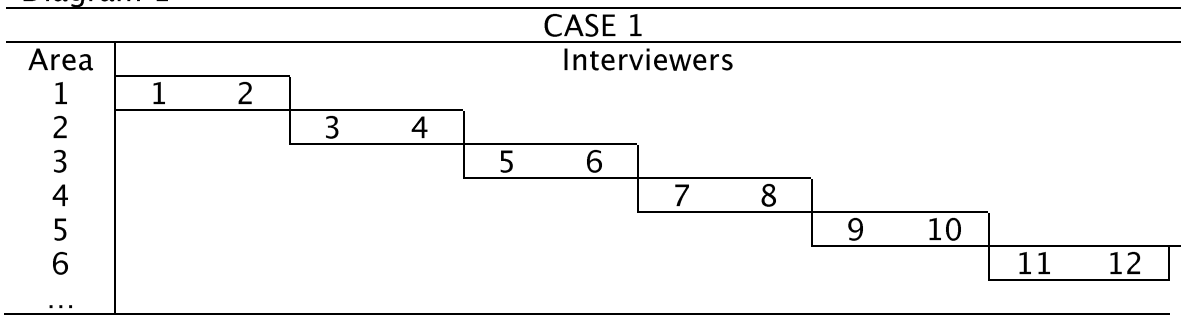

Diagram 2

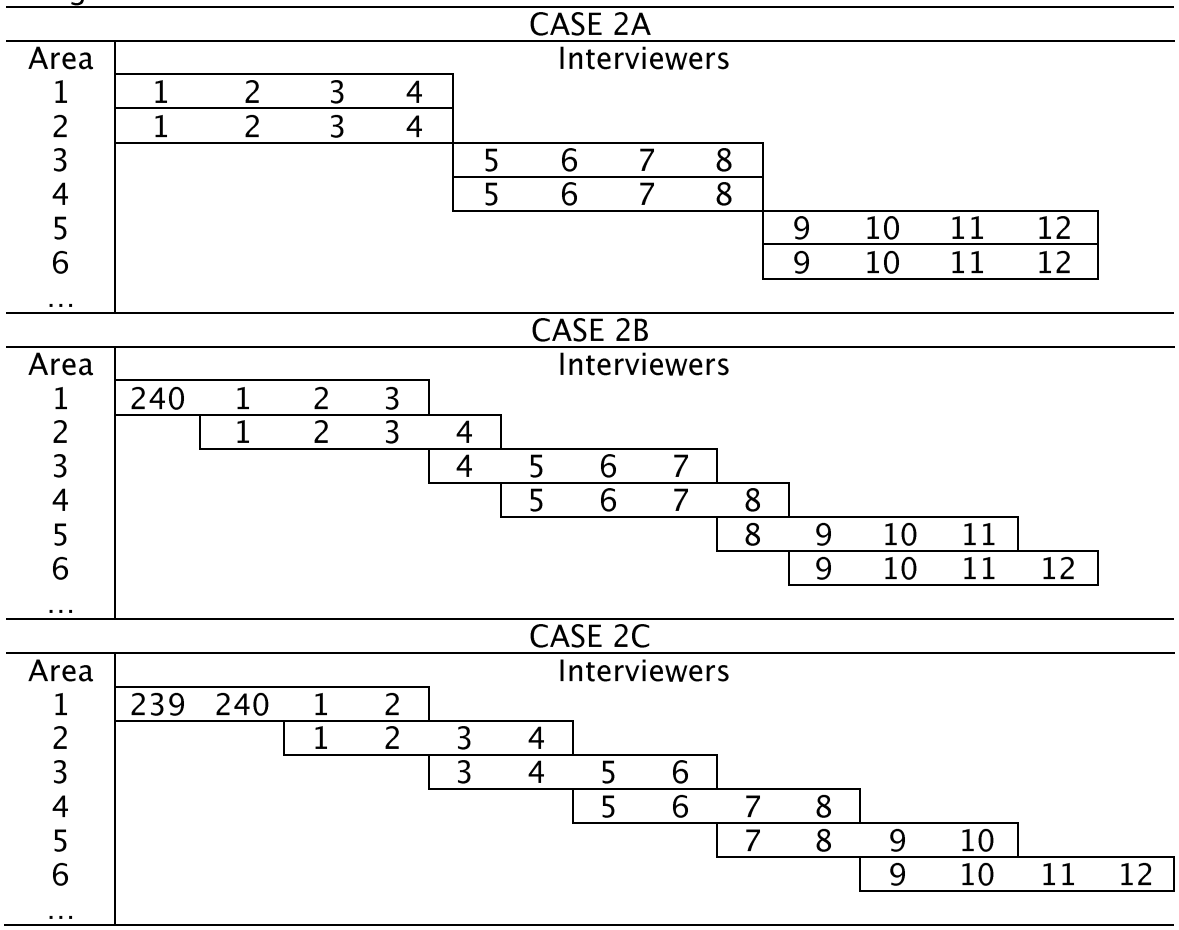

Diagram 3

CASE 1: $N^I=N^A$

Area	Interviewers
1	1
2	2
3	3
4	4
5	5
6	6
…	

CASE 2: $N^I=N^A$

Area	Interviewers
1	1 2
2	1 2
3	3 4
4	3 4
5	5 6
6	5 6
…	

CASE 3: $N^I=N^A$

Area	Interviewers
1	1 2 3
2	1 2 3
3	1 2 3
4	4 5 6
5	4 5 6
6	4 5 6
…	

CASE 4: $N^I=N^A$

Area	Interviewers
1	1 2 3 4
2	1 2 3 4
3	1 2 3 4
4	1 2 3 4
5	5 6 7 8
6	5 6 7 8
…	

CASE 5: $N^I=N^A$

Area	Interviewers
1	1 2 3 4 5
2	1 2 3 4 5
3	1 2 3 4 5
4	1 2 3 4 5
5	1 2 3 4 5
6	6 7 8 9 10
…	

CASE 6: $N^I=N^A$

Area	Interviewers

1	1	2	3	4	5	6	
2	1	2	3	4	5	6	
3	1	2	3	4	5	6	
4	1	2	3	4	5	6	
5	1	2	3	4	5	6	
6	1	2	3	4	5	6	
…							