UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

School of Social Sciences

Division of Social Statistics

Investigating the Performance of Multilevel Cross-classified and Multiple Membership Logistic Models: With Applications to Interviewer Effects on Nonresponse

by

Rebecca Vassallo

Thesis for the degree of Doctor of Philosophy

March 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

SOCIAL STATISTICS

Thesis for the degree of Doctor of Philosophy

INVESTIGATING THE PERFORMANCE OF MULTILEVEL CROSS-CLASSIFIED AND MULTIPLE MEMBERSHIP LOGISTIC MODELS: WITH APPLICATIONS TO INTERVIEWER EFFECTS ON NONRESPONSE

Rebecca Vassallo

This thesis focuses on the modelling of interviewer effects on nonresponse using cross-classified and multiple membership multilevel logistic models, and investigates the properties of such models under various survey conditions.

The first paper reviews the use of cross-classified and multiple membership models to account for both interviewer and area effects and for various wave interviewers. An extension to incorporate both wave interviewer effects and area effects is presented. The mathematical details, assumptions and limitations of the models are considered. The different models conceptualised are then fitted to a dataset. This application extends the focus of the first paper from simply a methodological one to an applied study with substantive research questions. The study aims to identify interviewer characteristics that influence nonresponse behaviour, assess the relative importance of previous and current wave interviewers on current wave nonresponse, and explore whether respondents react favourably to interviewers with similar characteristics.

The second and third papers investigate the properties of cross-classified and multiple membership multilevel models respectively under various survey conditions. The second study looks at the effects of different interviewer case assignment schemes, total sample sizes, group sizes (interviewer caseload), number of groups (number of interviewers), overall rates of response, and the variance partitioning coefficient on the properties of the estimators and the power of the Wald test. The study aims to provide practical recommendations for future study designs by identifying the smallest total sample size, interviewer pool, and the most geographically-restrictive and

cost-effective interviewer case allocation required to adequately distinguish between area and interviewer effects. The third paper includes a sensitivity analysis which looks at how accurately the Deviance Information Criterion identifies the best weighting scheme for different true multiple membership weights, interview allocation profiles, and total sample sizes. This sensitivity analysis indicates how well the relative importance of the previous and current wave interviewers can be estimated in multiple membership models under different survey conditions. Moreover the quality of parameter estimates under models with correctly specified weights, models with incorrectly specified weights, and models with weights based on the Deviance Information Criterion are also investigated.

Contents

List of tables

List of figures

DECLARATION OF AUTHORSHIP

I, Rebecca Vassallo, declare that the thesis entitled

'Investigating the Performance of Multilevel Cross-classified and Multiple Membership Logistic Models: With Applications to Interviewer Effects on Nonresponse'

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- part of this work has been published as: Vassallo, Rebecca, Durrant, Gabriele B., Smith, Peter W.F. & Goldstein, Harvey (2014). Interviewer effects on nonresponse propensity in longitudinal surveys: a multilevel modeling approach. Journal of the Royal Statistical Society Series A (early view online publication - doi:10.1111/rssa.12049).

Signed: ……..

Date:……

Acknowledgements

This PhD would not have been possible without the financial support of primarily the University of Southampton School of Social Sciences Teaching Studentship and also the PhD Studentship from the UK Economic and Research Council (ES/1026258/1). I am grateful for the opportunity the School has given me and for the wealth of knowledge I have obtained throughout my four years at this university.

I would like to express my heartfelt gratitude to my supervisors, Dr Gabriele B. Durrant and Professor Peter W. F. Smith, for their helpful comments, guidance, patience, and presence throughout the process of this dissertation. I am grateful to them for being a constant and ongoing point of reference and support in spite of their busy schedules, especially during critical times. They have been inspirational models of rigour, dedication and ethical professionalism.

I would like to thank my lecturers from the Department of Social Policy at the University of Malta, who encouraged and guided me at the start of my academic studies. I am particularly grateful to Mr Edgar Galea Curmi for instilling in me an avid interest in research and for going well beyond his obligations and being a mentor to me throughout the years. I would also like to mention Dr Charles Pace for his dedication to the Department and to the pastoral care of students, and for his encouragement and belief in me.

I am grateful to my upgrade examiner Dr James J. Brown for his useful comments and suggestions on the upgrade version of the thesis, and to my external advisor Professor Harvey Goldstein for his interest in and feedback on my work. I would like to thank the Centre for Multilevel Modelling at the University of Bristol for providing a platform for forwarding queries and problems regarding MLwiN and runmlwin as well as providing immediate replies to my posts. I would also like to thank the support services associated with the IRIDIS High Performance Computing Facility cluster at the University of Southampton. Special thanks go to Mr David F. Grain and Dr Oz G. Parchment

for patiently helping me set up a system that could work for my particular project and for providing technical support whenever problems arose.

I am grateful to my friends and fellow PhD students for their emotional support, care and companionship.

Finally my thanks go to my family, who have always been a pillar of strength to me. I thank them for their unconditional love, moral support, encouragement and patience throughout these years, especially during the last stages of the PhD.

List of Acronyms

Akaike Information Criterion (AIC)

Confidence Interval (CI)

Deviance Information Criterion (DIC)

Family and Children Study (FACS)

Intra-cluster Correlation (ICC)

Iterative Generalised Least Squares (IGLS)

National Centre for Social Research (NatCen)

Marginal Quasi-likelihood (MQL)

Markov Chain Monte Carlo (MCMC)

Multiple Membership (MM)

Multiple Interviewer Membership (MIM)

Multilevel Multiple Membership Multiple Classification (MMMC)

Penalized Quasi-likelihood (PQL)

Primary Sampling Unit (PSU)

Variance Partitioning Coefficient (VPC)

I. Introduction

I.1. Research Purpose

The aim of this thesis is to investigate the performance of cross-classified and multiple membership multilevel logistic models, with applications to interviewer effects on nonresponse.

I.2. Outline of the Thesis

The Introduction chapter outlines the aims and significance of the study and the rationale behind the choice of the three papers. It also includes a brief overview of the concept of paradata, interviewer effects on nonresponse and multilevel modelling. Next, the three papers are presented separately. Each paper includes an introduction, literature review, data and methodology, results, and a discussion and conclusion section. The Conclusion chapter presents an overview of the study findings in the context of previous studies, and discusses study limitations and suggestions for further work as well as the implications of the results for survey administration and survey data analysis.

I.3. Aims and Significance of this Study

With a continuous rise in survey nonresponse rates, there is an increase in studies focusing on interviewer effects on nonresponse which aim to identify survey management strategies to reduce nonresponse at the design stage or during data collection. With interviewers playing an important role in gaining response (Groves & Couper, 1998; Hox & De Leeuw, 2002), such studies acknowledge the relevance of research in this area to guide policies in survey administration in order to minimise any negative interviewer effects (Sinibaldi et al., 2009; Durrant et al., 2010). This study aims to contribute to this body of literature, with more emphasis placed on the modelling of such data as well as

the investigation of the properties of estimators and the test statistic of such models under different survey conditions.

The first paper reviews the use of cross-classified multilevel models to account for both interviewer and area effects in cross-sectional surveys, these being cross-classified multilevel models. The paper also reviews the use of both cross-classified and multiple membership multilevel models to account for current and previous wave interviewers on wave nonresponse in longitudinal surveys. Multilevel multiple membership multiple classification (MMMC) models and cross-classified models including three higher-level classifications are considered to be an extension of these models to incorporate both the area and the two interviewer effects. The two modelling possibilities have different underlying assumptions, which are discussed and used to analyse a dataset. This application highlights possible estimation problems in the use of these models for realistic data structures, such as when interviewers work in a very limited number of areas and when only a percentage of cases experience interviewer changes across subsequent waves.

This paper also has an applied focus, with various substantive questions being addressed. Often in-depth information on interviewers is lacking in studies analysing interviewer effects (Hill & Willis, 2001; Watson, 2003; Nicoletti & Peracchi, 2005), limiting the analyses to random effects or very limited demographic variables. In this study the rich data available – the Family and Children Study linked to administrative data and interviewer data from the NatCen (National Centre for Social Research) 2009 interviewer survey – makes the exploration of the above-mentioned objectives possible. The study aims to identify interviewer characteristics that are associated with nonresponse behaviour, assess the relative importance of previous and current wave interviewers on current wave nonresponse, and explore whether respondents react favourably to interviewers with similar characteristics. Results may be used to inform interviewer selection, training, appraisal and case allocation. The impact of matching the sampled member and the interviewer on demographic or socio-economic characteristics on the sample member's likelihood of responding is explored. Any significant matching effects may suggest criteria for successful interviewer changes at later waves.

The second and third papers look at the properties of cross-classified and multiple membership multilevel models respectively under various survey conditions. The second study looks at the effects of different interviewer case assignment schemes, total sample sizes, group sizes, number of groups, overall rates of response, and variance partitioning coefficient on the properties of parameter estimators and the power of the Wald test. The properties of parameter estimators include the covariance of the two variance estimates, percentage relative bias, the mean squared error, the standard error and the confidence interval coverage. The study will also provide practical recommendations for future study designs by identifying the smallest total sample size, interviewer pool, and the most geographically restrictive and cost-effective interviewer case allocation required to adequately distinguish between area and interviewer effects. The third paper includes a sensitivity analysis which looks at how accurately the Deviance Information Criterion diagnostic identifies the best weighting scheme for different true multiple membership weights, proportions of the total sample which experience an interviewer change, total sample sizes and number of groups for typical values of overall probability and intra-cluster correlations. This sensitivity analysis will indicate how well the relative importance of the previous and current wave interviewers can be estimated in multiple membership models under different survey conditions. Moreover, the quality of parameter estimates under models with correctly specified weights, models with incorrectly specified weights, and models with weights selected on the basis of the Deviance Information Criterion are also investigated. Methodologically the two simulation papers highlight the survey conditions under which models perform well for this type of data structure. However, in spite of the focus on survey nonresponse for cases from various area sampling units assigned to specific interviewers, the results on the performance of cross-classified and multiple membership models can be extended to other applications with similar structures.

I.4. Background

I.4.1. Paradata

The term *paradata* was first coined by Mick Couper at the Joint Statistical Meeting in Dallas (1998), and it initially referred strictly to automated computer-generated survey process data for CAPI and CATI systems. Call record data, keystrokes and other audit trails fall under this initial restrictive definition of paradata. The analysis of such data was being advocated to assess the quality of the inputted data and to identify any patterns in contact and response. Since then, as Kreuter (2010) explains, this term has evolved to include data which is generated during the survey process – but which is not part of the survey questionnaire itself – and collected using various methods. So, for example, interviewer observations of the surrounding area neighbourhood and audio recordings of doorstep interactions are nowadays also considered to be paradata, and are seen as information to potentially improve survey practice. A simple interviewer identification code linking an interviewer to a specific respondent case is also very important paradata. This is the main paradata dealt with in this dissertation. The interviewer code allows for the clustering of cases within interviewers to be correctly accounted for, and also allows the variation in the outcome measure that is attributable to the interviewer to be quantified. Other linked data such as administrative data or data from other surveys, for example interviewer historical response rates and demographic information or interviewer survey data, are not typically considered as paradata but simply as auxiliary data (Kreuter, 2010).

I.4.2. Interviewer Effects on Nonresponse

The presence of significant interviewer effects on nonresponse has been confirmed in both cross-sectional (Blom et al., 2010; Durrant & Steele, 2009; Durrant et al., 2010) and longitudinal surveys (Campanelli & O'Muircheartaigh, 1999; Pickery & Loosveldt, 2002; Pickery et al., 2001; Haunberger, 2010). In survey data, the non-random allocation of interviewers across areas raises questions as to whether interviewer effects are simply higher-level effects,

more specifically area effects, on nonresponse. Some studies, ignoring area effects and accounting only for interviewer effects in a multilevel model, simply find evidence of significant higher-level effects (Pickery & Loosveldt, 2002; Haunberger, 2010; Blom et al., 2010). Other studies, such as that carried out by Durrant et al. (2010), attempt to disentangle interviewer and area effects by specifying a cross-classified multilevel model for multistage cluster sample design data. Cross-classified models are multilevel models which allow the higher-level variance to be segmented into different non-nested cross cutting levels. These studies generally seem to indicate that interviewer effects are more important than area effects. In fact, the results in Durrant et al. (2010) show a highly significant interviewer variance around three times the size of the primary sampling unit variance, which is only marginally significant. In typical survey settings, although interviewers may work in more than one sampling area, these areas will be limited to neighbouring areas. The extent to which the limited cross-classification between areas and interviewers in real survey data impacts the accuracy and precision of the estimates obtained from these studies is not yet known.

The best data for determining the nature of the higher-level random effect is data coming from an interpenetrated sample design, where each sampled case is randomly allocated to interviewers irrespective of the area provenance of the case. Such a design also eliminates the possibility of more difficult cases systematically being allocated to more experienced and better performing interviewers at later waves, with the consequence that significant interviewer effects simply reflect clusters of sampled cases of varying difficulty as assessed by the field administrators rather than true interviewer-level clustering. A quasi-randomisation of cases across interviewers was implemented at the second wave of the British Household Panel Study, where the sample cases available for assignment for a specific interviewer were restricted to a geographic pool of two to three primary sampling units. The randomisation restriction was motivated by field administration capabilities and survey costs. Campanelli and O'Muircheartaigh (1999) analyse the significance of area and interviewer random effects on household and individual nonresponse, refusal and non-contact for data using a cross-classified multilevel model. The multilevel structure of the data pertains to

households within a cross-classification of interviewers within primary sampling units, nested within geographic pools. The authors find no evidence of random effects at the primary sampling unit and the geographic pool levels, but they do find some indication of interviewer effects, albeit non-significant.

Some studies have been investigating the relative importance of interviewers across two waves in longitudinal studies. Pickery et al. (2001) analyse non-response conditional on contact at the second wave using a cross-classified logistic model, with the interviewer at the first wave and the interviewer at the second wave specified as independent effects at the higher level. The authors find the first wave interviewer variance to be significant, whilst the random effect for the second wave interviewer is not. This result might be emphasising the importance of a positive first encounter, suggesting that the best interviewers within a survey agency may be allocated exclusively to the first wave of each rotating panel survey. Alternatively, which interviewer remains significant when both interviewer random effects are included in the model as independent effects may simply be random, devoid of substantive meaning.

The recent work by Lynn et al. (2013) suggests using multiple membership models to investigate the relative importance of distinct interviewers from different waves. Multiple membership models are multilevel models which include a higher-level variance for only one classification structure but allow each case to be associated with more than one higher-level unit. Each case has a weighted average of the individual higher-level contributions. In comparison to a cross-classified model, a multiple membership model does not assume that the effect of an interviewer with a specific case at a particular wave is distinct and independent of its effect at a different wave. Alternatively, the multiple membership specification makes a distinction between cases that are allocated the same interviewer across all the waves considered and those experiencing an interviewer change. Cases allocated to the same interviewer across both waves will only be attributed one interviewer effect which is constant across waves. On the other hand, two interviewer effects will be associated with cases experiencing interviewer change. The overall interviewer effects for cases with interviewer change will

simply be a weighted average effect of the two distinct interviewer effects. Lynn et al. (2013) use the Deviance Information Criterion, a Bayesian measure of model fit which penalises for model complexity (Spiegelhalter et al., 2002), to select the best model weights. As yet there has not been any work carried out to investigate the sensitivity of this measure in identifying the best multiple membership weights.

Although Lynn et al. (2013) find no significant average difference in the response probability by interviewer combinations, or even simply by current interviewers, they find a differential effect in the influence of respondent age on the respondent's probability of refusal by interviewer combinations. Their analysis shows that for cases experiencing interviewer change, the more recent interviewer has the bigger influence on the propensity to respond. The current wave interviewer is found to be responsible for 65% of the unexplained random respondent age slope by interviewer effect, while the previous wave interviewer is responsible for 35% in the final model. This interpretation is based on the multiple membership weights included in the final model. However, this differential effect in the influence of respondent age on the response probability by interviewer combinations is no longer significant in this final model, as it is explained away by including fixed effects of interviewer characteristics, namely interviewer change and interviewer age.

The confirmation of significant interviewer effects on survey nonresponse itself does not provide any indication of the survey administration strategy required to reduce nonresponse. Consequently, most studies identifying significant interviewer effects seek to explain this random variation by including interviewer-level fixed effects in the model. The fixed effects considered in previous literature on nonresponse include interviewer continuity, demographic and socio-economic characteristics, experience, attitudes towards respondent persuasion, on-the-job skills, and behaviour and personality traits. The lack of consistent relationships between these interviewer fixed effects and nonresponse partly reflects the lack of detailed information on interviewers available in many of these studies (Hansen, 2006; Haunberger, 2009). Some exceptions are studies such as those by Durrant et al. (2010), Blom et al. (2010), and Sinibaldi et al. (2009). These contain

information about interviewer attitudes in relation to the effectiveness and appropriateness of persuasion of reluctant respondents, typical doorstep behaviour, time organisation and availability, detailed information on experience, and skills and strategies employed on the job among others. More detailed information on the relationship between specific interviewer characteristics and nonresponse will be reviewed in the first paper, which will also test various hypotheses using the detailed interviewer data available, including both administrative data and data from an interviewer survey.

I.4.3. Multilevel Models

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. Standard analytical techniques cannot be used for clustered data since the violation of the assumption of independence of observations results in underestimated standard errors and can therefore result in incorrect inference (Rasbash, 2006; Snijders et al., 1999). Both disaggregated and aggregated approaches can adjust the standard error estimates to account for the dependency in the clustered data. While in aggregated methods design variables are only implicitly accounted for by averaging the effect of other explanatory variables over the population distribution of these design variables, in disaggregated methods design variables may be treated as scientifically relevant and are explicitly incorporated into the model.

Multilevel modelling, which is a disaggregated approach, allows for an extension of the error term included in standard regression analysis to be able to adjust for such dependencies. This extension consists of the inclusion of a residual error term for each classification in the structure. Consequently, multilevel models allow the variation in the outcome variable to be partitioned into various sources, these being both individual and group sources. The percentage variation of the outcome variable explained by fixed effects – which

can be specified at either the individual or the cluster levels – can be obtained by calculating the decrease in the unexplained individual and cluster variance after the addition of these fixed effects. 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. Despite allowing for a detailed analysis of contextual effects through the inclusion of a higher-level random effect and contextual or aggregate fixed effects, multilevel models include 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 as in 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.

Historically, multilevel models were first used in educational research applications to account for pupils in classes within schools – a purely hierarchical structure. However, with time multilevel models have been applied to diverse areas of study – which include organisational, demographic, biological and geographical data – and more flexible model specifications have been developed for more complex structures (Goldstein, 2011). The advances in more flexible models have also been aided by recent developments in Bayesian computation and efficient and powerful computing (Browne et al., 2001). Besides purely hierarchical structures, multilevel models can also deal with data pertaining to two different non-hierarchical classifications (cross-classified), as well as data where there is one classification but where individual cases may be associated with more than one higher-level unit (multiple memberships) (Fielding & Goldstein, 2006). The mathematical details of these multilevel models for the binary outcome case will be reviewed in the three thesis papers. The analysis of interviewer effects has become a popular application of these methods with various data structures conceptualised and a

range of multilevel model specifications employed to model such data (Von Sanden, 2004), as discussed in the next section.

The focus of this study is on multilevel models with a binary outcome. The general form of a logistic multilevel model for purely hierarchical data with two levels is:

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

The outcome represents 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 $\mathbf{X}_{ij}$, which refers to the mean log-odds for an individual pertaining to the reference categories of 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 parameter 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 $\mathbf{X}_{ij}$ represents the predictor variables which may be defined at the individual or cluster level (known as contextual effects). 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. However, estimation and inference procedures are usually robust to departures from this assumption (Fielding, & Goldstein, 2006, pp.36).

The model does not include an individual-level residual because the probability of the individual obtaining a value of 1 on the *y* variable is being modelled, rather than the value on a continuous outcome. The Variance Partitioning Coefficient (VPC) is a measure of the proportion of the total variance which arises due to cluster differences. For a simple multilevel model for which the variation between clusters varies at intercept only, the VPC is equal to the intra-cluster correlation, which represents the correlation of two response variable values selected at random from a particular cluster. To calculate the VPC, the individual-level variance can be assumed to be fixed at 3.29 using the threshold specification of a logistic model (Snijders & Bosker, 1999; Goldstein, 2011; Goldstein et al., 2002). Goldstein et al. (2002) explain that when the outcome variable is truly discrete and cannot be considered as underlying a continuous outcome (with a value of 1 representing some threshold in the continuous scale being exceeded), this threshold specification of a logistic model, which they call the latent variable approach, is not the best approach for calculating the VPC. Instead, the authors propose using model linearisation or simulation. However these two methods will not be used in this study since they do not offer a general summary for all the data, but VPC values for specific values of X. Moreover, the threshold specification method of calculating the VPC is widely used in applied studies using multilevel models for the analysis of discrete binary outcome data (Dundas et al., 2006; Johnell et al., 2004; Ferede, 2013).

Equation I.1 represents a random intercept model, where the group differences vary only in terms of the mean intercept. A possible extension to this model is the inclusion of a random term at the cluster-level for parameter coefficients, to allow the effect of an explanatory variable X to vary randomly across clusters. This extension would result in a random coefficient model. Another extension could be the inclusion of other levels for purely hierarchical data. Alternative random effects models may also be specified, such as cross-classified random effects or a multiple membership structure. Ignoring a level in a multilevel model (Tranmer & Steel, 2001; van den Noortage et al., 2005), or incorrectly specifying the random effects structure by ignoring a crossed factor (Luo & Kwow, 2009) or specifying a purely hierarchical model for

multiple membership data (Chung & Beretvas, 2012), can lead to biased estimation of the regression coefficients and the standard errors, and consequently result in wrong inference. It is therefore important to consider carefully the underlying data structure and to theorise correctly the structure underlying the data and model it adequately.

I.4.4. Multilevel Models for Interviewer Effects

Most survey data is hierarchical in nature, for example due to a multistage sampling procedure or, in the case of face-to-face and telephone surveys, the allocation of cases to interviewers. Clustering of nonresponse at the higher level may reflect various unmeasured factors, including an interviewer's demographic characteristics, skills, attitudes, behaviour and personality traits, or the cultural and socio-economic nature of the sampled area. Various analytical techniques for the analysis of interviewer effects have been developed (Von Sanden, 2004). These include the variance decomposition technique (Hansen et al., 1953), the correlation approach (Hansen et al., 1961), and the ANOVA linear model approach (Kish, 1962). Multilevel modelling is essentially a generalisation of these and has become a popular method of choice in research analysing interviewer effects (Hox & De Leeuw, 2002; Pickery & Loosveldt, 2002; Durrant & Steele, 2009; Blom et al., 2010; Durrant et al., 2010; Haunberger, 2010; Lynn et al., 2013). A disaggregated approach is the preferred method for this application since the area and interviewer clusters are of scientific interest in the study of nonresponse and may directly affect the outcome variable. Moreover, a disaggregated approach allows the impact of the population indicators (design variables) on the outcome variable survey nonresponse to be quantified.

One of the first applications of multilevel models for the analysis of interviewer effects goes back to 1985 in a study carried out by Anderson and Aitken (1985) aimed at measuring the variability in responses on consumer spending by interviewers. Advantages of multilevel models include their ability to account for different hierarchical structures of survey data and the treatment of clustering as an integral aspect of the analysis, rather than being

seen as a nuisance simply to be accounted for. In particular, multilevel models allow the investigation of substantive research questions that go beyond the scope of standard approaches, such as the possibility of analysing the amount of total variation attributable to interviewer effects. The use of cross-classified and multiple membership multilevel models to also control for area effects and the effects of various interviewers has also been attempted (Durrant et al., 2010; Lynn et al., 2013). These extended multilevel models allow the estimation of the relative impact of different wave interviewers on current wave nonresponse and the separation of higher-level effects into interviewer and area effects. Full mathematical details of these methods, the rationale for their use, and a discussion of any issues in their implementation will be presented in the first paper.

I.4.5. Monte Carlo Markov Chain Estimation Method

A range of estimation methods exist to estimate the parameters of multilevel models. Markov Chain Monte Carlo (MCMC) methods, using diffuse priors and the quasi-likelihood estimates as starting values (Goldstein & Rasbash, 1996; Goldstein, 2011), have been shown to produce improved estimates compared to first-order marginal quasi-likelihood (MQL) and second-order penalized quasi-likelihood (PQL) in terms of frequentist unbiasedness for multilevel logistic models (Browne, 1998; Browne & Draper, 2006). Moreover, Bayesian methods (of which MCMC methods are a sub-type) offer a general and more flexible approach to model complex data structures than likelihood-based methods (Browne, 2012). In this study MCMC estimation, with diffuse priors and the second-order PQL estimates as starting values as implemented in the MLwiN software, is being used. MLwiN is the most widely used specialist multilevel software in the UK (Fielding et al., 2006b). It allows large datasets to be imported and complex model structures to be fitted using both frequentist likelihood and MCMC estimation methods (Rasbash et al., 2012). MCMC estimation is being used simply to maximise likelihood for the unknown variance parameters. It has been shown that using the diffuse priors integrated in MLwiN gives similar estimates to maximum likelihood estimation (Browne,

2012, Chapter 6). As shown in Browne and Draper (2006), using the PQL estimates as starting values allows for a short burn-in period without the starting values having undue influence on the posterior distribution. The authors suggest that a burn-in period of 500 iterations should be sufficient. However, in practise 500 iterations will rarely be sufficient, especially for models with a sparse cross-classified or multiple membership structure. Consequently, it is best to attempt different burn-in lengths to identify the appropriate length of discarded iterations to avoid undue influence from the starting values (Gelman et al., 2004).

II. Interviewer Effects on Nonresponse Propensity in Longitudinal Surveys: A Multilevel Modelling Approach (Paper 1)

II.1. Introduction

II.1.1. Background

The decline in survey response rates, documented by De Leeuw and De Heer (2002), provides a strong motivation for investigating the causes and factors influencing nonresponse. Prominent among such studies are those analysing interviewer effects, which aim to reduce nonresponse at the design stage or during data collection. In their theoretical framework for household nonresponse, Groves and Couper (1998) identify five factors that influence the process of refusal, of which interviewer attributes and the interviewer-respondent interaction represent two out of only three factors which the survey agency has some control over. Studies focusing on interviewer effects reflect the understanding that interviewers play an important role in introducing the survey concept, engaging the respondent, addressing any queries, and ultimately gaining response (Groves & Couper, 1998; Hox & De Leeuw, 2002). They also acknowledge the possible influence the research agency can have in minimising negative interviewer effects through effective policies and management strategies (Sinibaldi et al., 2009; Durrant et al., 2010).

While a number of studies have confirmed the presence of significant interviewer effects on nonresponse in both cross-sectional (Durrant & Steele, 2009; Blom et al., 2010; Durrant et al., 2010) and longitudinal surveys (Campanelli & O'Muircheartaigh, 1999; Pickery et al., 2001; Pickery & Loosveldt, 2002; Haunberger, 2010), there has been little conclusive or consistent evidence concerning the interviewer attributes associated with higher response rates. This partly reflects the lack of detailed information on interviewers available in many of these studies.

Some attention has been given to the modelling of interviewer effects on nonresponse in sample surveys. For cross-sectional data, a multilevel modelling approach has been advocated (Hox & De Leeuw, 2002; Durrant & Steele, 2009; Blom et al., 2010; Durrant et al., 2010). A complicating factor when analysing interviewer effects is that 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. 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).

For longitudinal surveys, there are added complexities in the analysis of interviewer effects on nonresponse. Firstly, the same case may be allocated to different interviewers at different waves. Secondly, the response outcome for the same sampled person may vary across waves. Finally, there may be changes in the sample owing to, for example, changes in the eligibility criteria or booster samples aimed at restoring the sample representativeness after attrition. On the other hand, longitudinal data offers the advantage of obtaining information on both respondents and nonrespondents from previous waves when analysing cooperation at a later wave, which is often missing for cross-sectional data. A particular research interest, of importance for effective longitudinal survey designs, is the optimal interviewer allocation across waves for the same respondent. This decision must take into consideration the impact of interviewer change and the relative impact of distinct interviewers from previous waves on respondent cooperation at later waves.

There is, however, limited research that takes into consideration the modelling of interviewer effects in longitudinal surveys, taking account of the additional complexities of such surveys, in particular the influence of potentially more than one interviewer across waves. Pickery et al. (2001) propose a multilevel cross-classified logistic model, specifying the effects of the previous and current interviewers as independent effects at the higher level. A more recent working paper by Lynn et al. (2013) uses a multilevel multiple membership model (Goldstein, 2011), where the overall interviewer

effect is made up of a weighted combination of two wave interviewers. This model incorporates the effect of distinct interviewers associated with a particular case across different waves. However, the model does not differentiate the effect of a particular interviewer across different waves for cases with interviewer continuity. Consequently, the relative benefits of different modelling approaches for the analysis of interviewer effects in longitudinal surveys have not been considered. Also, there has been no study that included various wave interviewer effects and the area effect in one model.

II.1.2. Aims and Methods

This paper presents a multilevel modelling framework for the analysis of interviewer effects on wave nonresponse in longitudinal surveys and considers in particular two different multilevel modelling approaches: the multilevel cross-classified and the multiple membership models. The proposed modelling approach incorporates both interviewer and area effects accounting for the non-hierarchical structure and potential confounding due to a lack of an interpenetrated sample design arising from the non-random allocation of interviewers across areas.

The methods are illustrated using a dataset from the UK Family and Children Survey, with the focus on nonresponse at a later wave in the life of a longitudinal study. Both cross-classified and multiple membership specifications will be considered to account for wave 7 and wave 8 interviewers. The two most complex multilevel logistic models fitted for modelling nonresponse at wave 8 are a cross-classified multilevel logistic model with three independent random effects and a multilevel multiple membership multiple classification (MMMC) model.

In addition to the methodological considerations, the study also makes some substantive contributions. The study aims to identify interviewer characteristics that are associated with nonresponse behaviour, with particular focus on interviewer experience, and personality and skills traits indicators. The study also considers the relative importance of interviewers across two

waves. An additional aim of the study is to consider whether respondents react favourably to interviewers with similar characteristics. This study uses observational data, and not experimental data, and therefore caution must be exercised in interpreting results. Causation cannot be inferred wherever a correlation is identified. Controlling for alternative causes of observed interviewer effects provides some evidence, but no certainty, for the presence of causal effects.

The research may have various implications for survey practice and data analysis. The study provides guidance to survey researchers on how best to model the relative influence of several interviewers on nonresponse across waves. The results may inform decisions how best to allocate interviewers across waves and across cases. For example, to the extent that the first wave interviewer has the greatest impact on the response at subsequent waves, it may be best to allocate work from the first waves of various surveys to the best interviewers within a survey agency. The identification of significant interviewer socio-demographic characteristics, work history, personality traits and job attitudes may provide guidelines for more effective interviewer recruitment, training, appraisal and work allocation. Just as there is some evidence, albeit weak, that a change for an interviewer of a similar age may be beneficial in achieving response at a later wave from older sampled members (Lynn et al., 2013), significant interviewer-respondent matching effects may suggest other criteria by which to determine successful interviewer changes.

The remainder of the paper is structured as follows. The data section presents the example dataset that is used to demonstrate the implementation of the proposed methods. The methodology section outlines the multilevel modelling approaches proposed and considers their relative benefits. The application section describes the implementation of the different multilevel models to the example dataset. It also provides an interpretation of the results of the final model and reviews non-significant effects. The final section discusses implications for survey practice and presents recommendations for future research.

II.2. Data

II.2.1. Main Data Source

II.2.1.1. General Purpose and Study Design

The example dataset is the longitudinal Family and Children Study (FACS), which gathers information on the health and socio-economic status of households with children in the United Kingdom (Lyon et al., 2007). This dataset is used to assess and illustrate the use of sophisticated multilevel models to investigate interviewer effects on nonresponse. The FACS benefits from the availability of rich survey information for both respondents and nonrespondents from previous waves of the longitudinal study. The FACS started in 1999 with a narrow focus on low-income families with children and lone-parent households, was expanded in 2001 to be representative of all households with children, and continues to this date. The study has a two-stage sampling design (Department for Work and Pensions, n.d.). First, a sample of 150 primary sampling units (PSUs) stratified by region and a rural/urban indicator from a total of 2600 postcode sectors (each representing 3000 households on average) listed in the Child Benefit database were chosen with probability proportional to the number of child benefit records. Secondly, a systematic sample of one hundred households with a random start was chosen within each cluster, resulting in 15,000 households prior to any reductions arising from address invalidity, opt-outs and screening procedures.

II.2.1.2. Data Structure

For the example analysis the focus is on the nonresponse behaviour at the later stages of this longitudinal survey, here on the last two waves for which all relevant information was available. These are wave 7 and wave 8 conducted in 2005 and 2006 respectively. The initial dataset includes the wave 8 cases that had participated in wave 7. A wave number represents the number of survey episodes since its inception in 1999. As there are varying initial waves and numbers of interviews for different cases considered in the dataset, participation history variables are included as controls in the multilevel models.

A complete case analysis is carried out. Therefore, cases having missing values for interviewer variables used in the final model had to be dropped. Some descriptive statistics, looking at frequency distributions of the response outcome by potential explanatory variables, have been run for both the full and restricted dataset. These tables, included in Appendix A, do not show any differences in the dataset profiles. Similarly to Sinibaldi et al. (2009), unit nonresponse in the interviewer survey is controlled for by including variables from the interviewer administrative data available for all interviewers, such as gender, experience and grade. The final analysis sample, referred to as the restricted dataset due to the dropped cases with missing interviewer data, includes 5932 cases pertaining to 307 wave 7 interviewers, 275 wave 8 interviewers, and 150 PSUs.

For the analysis of nonresponse no distinction needs to be made between the individual and the household level, as the only eligible respondent is the mother in the household, or the father in the case of a single-male-parent household. Interviewers are not nested within areas, as one interviewer may work in more than one primary sampling unit, and cases within one primary sampling unit may be assigned to more than one interviewer. For wave 8 there are no PSUs in which only one interviewer was allocated work, and approximately 82% of interviewers were allocated households from at least two different PSUs. This results in a cross-classification of interviewers by PSUs. About 68.3% of cases changed their interviewer between waves 7 and 8, such that 73.1% of wave 8 interviewers had cases associated with different interviewers across the two waves. Full details of the distribution of the number of interviewers per area and the number of areas per interviewer are presented in Appendix L. A change in interviewer allocation may arise because of a move of the household, changes in the interviewer's responsibility and workload, or the possibility that the interviewer may have stopped working for the survey agency.

Similarly to most other studies, with the exception of some experimental studies – such as Campanelli and O'Muircheartaigh (1999), Schnell and Kreuter (2005) and Lynn et al. (2013) – the allocation of cases to interviewers cannot be guaranteed to be completely random. However,

considering the large amount of interviewer variables available for this study, it is likely that any non-random criteria used to determine work allocation has been controlled for.

II.2.1.3. Response Variable

Cooperation or refusal with regard to the main face-to-face survey interview is being investigated. The main outcome of interest is whether or not a person responded to wave 8, conditioning on response to wave 7. This is to allow detailed information on both the respondents and the nonrespondents to wave 8 to be obtained from the previous wave. The analysis is conditional on contact being made with the household, similar to the work by Watson and Wooden (2006), Blom et al. (2010), and Durrant et al. (2010). Nonresponse may arise from either non-contact or refusal at a later stage. As shown in various studies, including the studies by Pickery and Loosveldt (2002) and Durrant and Steele (2009), the processes and predictors of non-contact and refusal are different, and therefore a distinction is required. Only 2.6% of cases in wave 8 resulted in non-contacts, compared to 8.5% of cases resulting in refusals or unproductive interviews. The small number of non-contact cases and the added model complexity resulting from a distinction between these two outcomes led to the decision to analyse response conditional on contact. Moreover, the study of non-contact is more important for cross-sectional studies than for later waves of longitudinal studies.

The distribution of interviewer-level refusal rates is highly positively skewed, with 90% of the 335 wave 8 interviewers showing refusal rates of less than 20%, with a median of 7.7%, an average of 9.5% and a standard deviation of 9.9%. While 79 interviewers were successful in achieving cooperation for all of their contacted cases, one interviewer obtained only refusals. On the other hand, there is less variation for PSU-level refusal rates with a standard deviation of 5.5%, and the distribution shows closer, though not perfect, adherence to the normality assumption. The discrepancy between the median and the mean PSU-level refusal rate is lower than that for the interviewer-level rates, with values of 8.7% and 9.1% respectively. Four PSUs include only

cooperating households and the worst outcome pertains to one PSU achieving only 60% cooperation rate.

II.2.1.4. Household and Area Variables

The only area variables included in the FACS are a region variable and seven neighbourhood perception variables. The latter variables provide an individual-based assessment of how prevalent neighbourhood problems are in the area one lives in. While household-level variables are considered, the main purpose of their inclusion in the model will be as control variables, in the exploration of significant interviewer-respondent matching effects or cross-level interactions, and to offer some control for possible area confounding (to the extent that contextual effects are area averages of household-level characteristics). The identification of household- or individual-level predictors of nonresponse, which generally provide a profile of weighting variables for post-survey adjustment, is of restricted use within the context of this study's aims – to identify strategies which may help reduce refusals. At best, throughout the life of a longitudinal survey these variables may aid the identification of cases which are more likely to refuse. The allocation of such difficult cases to interviewers with the best track record of response rates may have a positive effect on the overall cooperation rate.

II.2.2. Additional Data Sources

A key advantage of the current study is that detailed information on interviewers is available from both administrative data on interviewers collected annually and from a survey of interviewers conducted in 2008, and both data sources have been linked to the survey data. The only other piece of work which makes use of this interviewer survey data is the study by Sinibaldi et al. (2009), which investigates nonresponse for all cross-sectional surveys conducted by NatCen in the period covering December 2007 to December 2008. Therefore, this is the first use of this data for the analysis of nonresponse in a longitudinal study.

II.2.2.1. Participation History Variables

The available participation history information includes a variable specifying the wave at which each case was first introduced to the sample, and seven dummy variables (for waves 1-7) indicating whether the case was interviewed at that particular wave. A variable providing a measure of the number of times each case was interviewed can be computed from these dummy variables.

II.2.2.2. Interviewer-level Variables

Information on interviewers firstly comes from administrative data collected by the survey agency NatCen on an annual basis. The administrative data includes the identification code of the interviewer allocated to each case for all waves, some demographic information, the interviewer grade as at mid-February 2008, corresponding to the interviewer survey date, and the years of experience within NatCen as at the beginning of September 2005. Interviewers are identified by the same identification code across waves, allowing an indicator of interviewer change between wave 7 and wave 8 to be constructed. The demographic information is restricted to the gender and age of the interviewer. Despite the lack of time correspondence between the experience and grade variables, changes in interviewer experience are constant across time and therefore this variable still indicates differences in experience between interviewers at any point in time. A key advantage of the interviewer administrative data is that the information is available for all interviewers, regardless of whether they responded to the interviewer survey or not. Therefore, interviewer administrative variables can be used in the multilevel models to control for unit-nonresponse to the interviewer survey.

Secondly, the interviewer survey, a postal self-completion survey administered in May 2008 addressed to all interviewers who had worked for NatCen at some point since the start of 2006, provides rich data on interviewing experience, job expectations and appraisal, flexibility in working hours, personality traits, inter-personal skills, and views on the persuasion of reluctant sample members. The survey had a comparatively high response rate

with just over 80% of eligible interviewers completing the interviewer survey (Sinibaldi et al., 2009).

The survey includes a 15-item personality assessment tool, devised for the German Socio-Economic Panel Study (Benet-Martínez & John, 1998), composed of three items for each of the Big Five personality traits. The interviewer makes a self-assessment for each item on a seven-point scale, where 1 indicates that the statement does not apply at all to oneself and 7 indicates that the item applies perfectly. This short test has undergone pre-tests and cross-validations with other reputable personality tools. Consequently, the composite measure of personality traits used in this paper follows the data structure validated across various studies, that is, 5 sets of 3 items. Therefore, no factor analysis is carried out, but the items pertaining to the specific personality trait – as prescribed in the tool documentation – are aggregated appropriately. Neuroticism refers to the tendency to experience negative emotions including stress, depression and anger, and the inability to cope well with difficult situations and to control impulses. Extroversion is characterised by a willingness to engage with others, take up challenges and be at the centre of attention. Conscientiousness may be described as a predisposition for self-discipline, respect for rules and duties, and a preference for planned rather than spontaneous activities. Agreeableness relates to a disposition to be compassionate, cooperative and altruistic. Openness is related to curiosity, appreciation of original and creative work, and a keen interest in new experiences.

The survey also includes 52 items which relate to skills which may help interviewers attain contact and cooperation. The International Personality Item Tool was used to inform the choice of the skills items to be included in this interviewer survey. Unlike the personality traits, these skills relate to specific behaviours that can be learnt. Similarly to the personality items, interviewers assessed themselves on these skills by rating how well each statement applied to their abilities from a scale of 1 to 7. In a similar manner to the study by Sinibaldi et al. (2009), only the 35 items related to cooperation are considered in this analysis. These items have been grouped into ten components – reading others, connectedness, verbal communication, nonverbal communication,

small talk, adaptability, ability to conform, assertiveness, deliberation, and emotional resilience – using principal component analysis.

Eight items gauge the interviewer's attitude towards the persuasion of reluctant respondents on a four-point scale. The items consider the interviewer's views on the effectiveness and acceptability of persuasion, the reliability of answers obtained after persuasion efforts, and the voluntary nature of survey cooperation.

Other items consider the work history of the interviewer. These include indicators for interviewing work experience with an agency other than NatCen, non-survey interviewing, other survey interviewing such as marketing or phone interviewing, work involving interaction with the general public, work involving cold calling and work requiring persuasion skills. Information about the current work profile is also included. The survey enquires whether the interviewer is also occupied by any other non-survey paid job, and also whether this other job is carried out at fixed times. Time availability for the NatCen work is considered in detail. Interviewers were asked whether they were unwilling or unable to do survey work during various time slots of the week.

Five items assess the importance an interviewer places on different aspects of a job on a three-item scale. The job characteristics considered include flexibility in working hours and workload, autonomy, financial remuneration and social interaction. The current NatCen job is appraised through various items requesting an assessment – on a scale from 1 to 4, from very satisfied to not satisfied at all – of various specific aspects of the job, including the job characteristics outlined for the personal job priorities items, and one overall satisfaction item for the job in general. Another set of job satisfaction factors are the items considering the level of support provided by various members of the organisations with leadership positions. The support from various key figures in the organisation, including the supervisor, team leader and area manager, is evaluated on a four-point scale.

II.2.3. Time Discrepancies in Data Sources

It should be noted that in analysing refusal for wave 8 of the FACS (2006), the interviewer variables obtained from the interviewer survey (2008) represent the interviewer characteristics approximately two years later. However, this discrepancy is not believed to have a significant impact on the results since most information from the interviewer survey used here, e.g. information on behaviours and personality traits, is assumed to be relatively stable over time (Sinibaldi et al., 2009). The time difference may also be beneficial to the extent that the actual performance on the job in terms of nonresponse is not directly driving the answer choice in the interviewer survey. Therefore, for example, replies to questions on views on respondent persuasion and questions on job satisfaction will not be directly affected by the interview experiences for which nonresponse is being analysed.

II.3. Methodology

II.3.1. Accounting for Area and Multiple Interviewer Effects across Waves: Cross-classified and Multiple Membership Models

A complicating factor when analysing interviewer effects is their potential confounding with areas (PSUs). For many face-to-face surveys an interviewer will work almost exclusively in a limited geographic area. Therefore, variation in the probability of refusal by interviewer may simply reflect area differences in the geographic propensity to cooperate in survey requests. Very few studies exist that have been able to use an interpenetrated sample design (Campanelli & O'Muircheartaigh, 1999; Schnell & Kreuter, 2005), where interviewers are randomly allocated to households, at least within a wider geographic pool of PSUs, enabling, to some extent, a separation of interviewer and PSU effects. A fully random allocation of interviewers to households for face-to-face surveys would be very costly and therefore practically very difficult. As a consequence, often such potential confounding is ignored in the analysis (Pickery &

Loosveldt, 2004). However, in some studies, as is the case here, a complete confounding of interviewers and areas is avoided, since interviewers and areas are partially interpenetrated (Von Sanden, 2004; Durrant et al., 2010). This means that interviewers are not fully nested within areas, as one interviewer may work in more than one PSU, and cases in one PSU may be designated to more than one interviewer. With a data structure showing partial interpenetration, a multilevel cross-classified model specification which considers both interviewer and area random terms can allow for a distinction between interviewer and area effects (Goldstein, 2011, Chapter 12).

The easiest way of accounting for the influence of interviewers in a longitudinal survey is to consider only the current wave interviewer. However, as shown in Goldstein (2011, Chapter 13), assigning a case to just one level-2 unit when in actual fact the case has multiple memberships – i.e. it belongs to more than one higher level unit (in this case interviewers) – will lead to an underestimation of the higher-level variance. It may be hypothesized that more than one interviewer, and potentially all interviewers associated with each case, have an influence on the nonresponse outcome. In this paper two different ways of specifying the various wave interviewer random effects are considered. One is to specify these effects as cross-classified (Goldstein, 1994), as was done in the study by Pickery et al. (2001), therefore assuming that each interviewer has a separate effect for each wave, and that the interviewer effects for each wave are independent. However, in the context of (at least some) interviewer continuity across waves, the tenability of this assumption is questionable. An alternative approach involves the use of a multilevel multiple membership model (Goldstein, 2011). The only application of multiple membership models in the analysis of interviewer effects on nonresponse is the paper by Lynn et al. (2013). The multiple membership specification takes account of all the distinct interviewers each case was allocated to across the various waves considered. Multiple membership models allow the effect of all distinct interviewers associated with a specific case to be incorporated in the model by attributing a weight to each interviewer effect; together these sum to a weight of 1, such that the estimated interviewer effect becomes a weighted average of all interviewers. These weights represent the

relative effect of each distinct interviewer. For multiple membership models interviewer effects are not wave specific.

Controlling for both complicating factors simultaneously – i.e. the confounding of area and interviewer effects and the influence of multiple interviewers per household across waves – leads to two possible specifications. Under the assumption of independent interviewer effects, a cross-classified model with various distinct random effects – one area effect and an interviewer effect for every wave – is obtained. Under the multiple interviewer membership assumption, a MMMC model specifying an area random effect cross-classified with the interviewer multiple membership is obtained. MMMC models allow the integration of cross-classified and multiple membership random terms specified at the same higher-level (Browne et al., 2001).

II.3.2. Model Specifications

Let $y_{i(\mathbf{j}s)}$ denote the dependent binary variable of interest, indicating whether individual *i*, interviewed by interviewers $\mathbf{j} = (j_1, \ldots, j_n)$ in waves $k = 1, \ldots, n$ and living in PSU *s*, refused to participate at wave *n*. Contact at wave *n* and response at wave *n*–1 are assumed. The general forms of the two most comprehensive multilevel logistic model specifications considered to model nonresponse at a particular wave *n* are presented below. While both models include a cross-classified area effect, the first model considers the various wave interviewers as cross-classified, while the second model considers a multiple membership for the interviewer allocation.

The general form of the cross-classified multilevel logistic model is:

$$\log\left(\frac{\pi_{i(\mathbf{j}s)}}{1-\pi_{i(\mathbf{j}s)}}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_{i(\mathbf{j}s)} + \sum_{k=1}^{n} u_{j_k}^{k} + v_s \qquad \text{(II.1)}$$

where $\pi_{i(\mathbf{j}s)} = \Pr(y_{i(\mathbf{j}s)} = 1)$ is the probability of individual *i* refusing to participate at wave *n*. The parameter β_0 represents the overall intercept in the linear relationship between the log-odds of refusal and predictor variables specified

in the model, $\mathbf{X}_{i(\mathrm{j}s)}$. The vector $\boldsymbol{\beta}_1$ contains the coefficients for the explanatory variables in the model. The parameters $u_{j_1}^1,\ldots,u_{j_n}^n$ and v_s represent the random effects for each wave interviewers $j_1,\ldots,j_n$, and the individual's area of residence respectively, which are assumed to follow a normal distribution with variances $\sigma_{\mathrm{u1}}^2 \ldots \sigma_{\mathrm{un}}^2$ and σ_v^2. The model includes an independent random effect for each wave interviewer considered.

The general form of the MMMC approach – the MMMC multilevel logistic model – is:

$$\log\left(\frac{\pi_{i(\mathrm{j}s)}}{1-\pi_{i(\mathrm{j}s)}}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_{i(\mathrm{j}s)} + \sum_{k=1}^{n} \mathrm{w}_{ij_k}\, u_{j_k} + v_s, \qquad \sum_{k=1}^{n} \mathrm{w}_{ij_k} = 1. \qquad \text{(II.2)}$$

The outcome variable and the fixed effects for the MMMC model have the same meaning and interpretation as those for the cross-classified model. The MMMC model includes one overall interviewer random effect u_{j} which, after combining weights for the same interviewer, is a weighted average of *m* distinct interviewers associated with a case in the *n* waves considered. After combining weights, cases allocated to the same interviewer across all waves have a weight of 1 for wave *n* and a weight of 0 for all previous waves, while cases experiencing an interviewer change have at least two non-zero weights summing to 1. The model formulation and the way the models are set up in MLwiN requires this combined weight specification for cases associated with only one higher-level unit (Browne, 2012). These combined weights are represented by the term $\mathrm{w}_{\mathrm{ijk}}$. The number of non-zero weights is thus equal to the number of distinct interviewers *m* associated with that particular case. Therefore each case may have from 1 to *n* distinct interviewers. Whilst there is only one common distribution for all interviewer effects from various waves, each individual has *m* distinct interviewer effects. In contrast, for the cross-classified multilevel model there are *n* unique interviewer effect distributions, one for each wave considered. The MMMC model also includes an area effect v_s cross-classified with the interviewer effect. Both the interviewer effect u_{j} and the individual's area residence effect v_s are assumed to follow independent normal distributions with variances σ_u^2 and σ_v^2.

The choice of which weights to apply can be empirically or theoretically based. There are three different justifiable theoretical arguments for the choice of weights. One possible argument is to allocate weights that are proportional to the number of waves the interviewer was allocated the specific case, reflecting the amount of time covered by each interviewer. Alternatively, one may argue that while the decision to take part in the current wave will be somewhat influenced by the experience in the prior waves, the current wave experience has a greater impact – since the current wave interviewer has the possibility to actively interact and address any hesitancies – and should therefore be given a greater weight. Another possible argument is that the current wave response decision is based on the experience at previous waves. It is plausible that the commitment and engagement obtained, or alternatively, the frustration and disassociation caused by the previous interviewer, has a greater impact than the current interviewer persuasion skills. A different theoretical argument may be valid for different surveys or for different parts of the longitudinal study. However, for all three options there is still some arbitrariness in the choice of weights. The weight profile can vary by both the number of interviewers and the sequence of interviewer changes. Alternatively, the choice of weights may be guided by an empirical assessment, as proposed in Goldstein (2011) and advocated in Lynn et al. (2013), using the Deviance Information Criterion (DIC) value for alternate models including various weight specifications. The DIC is a Bayesian measure of model fit which penalizes for model complexity, therefore allowing non-nested models to be compared (Spiegelhalter et al., 2002). As explained in the paper by Spiegelhalter et al. (2002), similarly to the Akaike Information Criterion (AIC) when comparing DIC values a model with a DIC value of 1-2 points lower than the current best model should be given consideration, while a model with a DIC value of at least 3 points lower than the current best model definitely has a better fit.

The weighted average of interviewer effects in the multiple membership model acts to dilute the higher-level effect (Fielding & Goldstein, 2006). For example, an individual allocated two different interviewers, each with an interviewer effect one standard deviation above the mean, will have an overall interviewer effect of one standard deviation above the mean when allocating equal weights in a multiple membership model. In contrast, in a cross-

classified model, this individual would have a higher-level effect of two standard deviations above the mean, with each interviewer contributing equally to this higher-level effect. The greatest dilution of these effects in the multiple membership model is observed when allocating equal weights to the higher-level effects. For models accounting for the previous and current interviewer effects the between-interviewer variance is equal to $(w_{ij_p}{}^2 + w_{ij_c}{}^2)\sigma_u^2$. This variance is at its lowest value when the interviewers are given equal weights.

The multiple membership model and this between-interviewer variance calculation are based on the assumption that the random effects for all interviewers, and pairings of interviewers, are mutually independent (Steele at al., 2013). In real terms this assumption means that, conditional on the explanatory variables included in the model, the choice of interviewer allocation at the previous and current wave is not based on the propensity of the individual to respond. To the extent that the allocated interviewers are chosen specifically on the basis of their performance record, and matched with the case difficulty, a correlation between the two interviewers would be present. If the most successful interviewers are targeted towards areas with high nonresponse rates a positive correlation between interviewer effects at different waves would be present, as households in difficult areas are allocated a pair of the best interviewers across both waves. Failing to account for this positive correlation would result in an underestimation of the between-interviewer variance. On the other hand, if better interviewers are allocated at the subsequent wave, following a poor outcome at the previous wave, a specific respondent request for a change in interviewer or to compensate for a change in interviewer due to interviewer attrition, then a negative correlation would be expected. Ignoring this negative correlation between the interviewer effects would result in an overestimation of the between-interviewer variance. Alternatively, if the choice of a new interviewer lacks strategy, and is simply based on interviewer availability and area proximity, then no correlation would be expected.

The Variance Partitioning Coefficient (VPC) is calculated to measure the proportion of total variance attributable to differences between interviewers. All the models specified allow the variation between interviewers to vary only

at the intercept. Consequently, the VPC is equal to the intra-interviewer correlation, which represents the correlation of two response propensity values selected at random from a particular interviewer. For any random intercept model, the intra-interviewer correlation is

$$\text{VPC} = \frac{\text{between interviewer variation}}{\text{total variation}} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2+\sigma_v^2} \qquad \text{(II. 3)}$$

where σ_u^2 is the interviewer-level variance and σ_e^2, the individual-level variance, is specified as 3.29 using the threshold specification of a logistic model (Snijders & Bosker, 1999; Goldstein, 2011; Goldstein et al., 2002).

II.3.3. Model Estimation and Modelling Strategy

Both cross-classified and multiple membership models can be estimated using Markov Chain Monte Carlo (MCMC) methods, using the quasi-likelihood estimates as starting values (Goldstein & Rasbash, 1996; Goldstein, 2011), as implemented in the MLwiN software (Browne, 2004). A forward selection strategy is used. The first step is the specification of the random effects structure, excluding any covariates in the model. To better understand the random terms and to identify whether the more complicated models are an improvement on simpler models, multilevel models including only one random effect are fitted first, and then more random terms added until the two most complicated models are obtained. The DIC diagnostic is used for model comparison, with a smaller DIC indicating a better fit. Significance testing for random effects can be based on the Wald test, despite the fact that a Wald test for random parameters is only approximate since variance estimates do not have normal sampling distributions under the null hypothesis, even asymptotically. Since variances cannot be negative, the alternative hypothesis is one-sided; therefore, one-sided *p*-values are used (Snijders & Bosker, 1999). Once an appropriate random structure specification is identified, groups of explanatory variables – participation history, household, area and interviewer variables – are considered for inclusion in a logical order.

For the multiple membership specification, the weights corresponding to the model showing the lowest DIC value without fixed effects are assumed for all further modelling steps, similar to the strategy used by Lynn et al. (2013). The procedure of choosing the optimal weight is then repeated once the final model, including all significant fixed effects, is identified.

II.4. Application

II.4.1. Hypothesised Relationships between Interviewer Variables and Response

The hypothesised relationship between the available interviewer variables in the FACS dataset and the response variable will be presented below. Any literature informing the hypothesised relationship is also highlighted.

II.4.1.1. Interviewer Continuity

The dummy variable indicating a change of interviewer between waves is expected to have a positive relationship with refusals. Studies considering the effect of interviewer changes across waves of a longitudinal study using observational data seem to confirm the common belief in survey administration that interviewer continuity is conducive to higher contact and response rates (Schatteman, 2000; Haunberger, 2010; Watson & Wooden, 2006). This common belief is mainly based on tradition rather than rigorous statistical evidence. This practice has implications both in terms of the optimal use of resources, particularly interviewers and finance, and on survey quality.

Campanelli and O'Muircheartaigh (2002) argue that in non-experimental studies of interviewer continuity effects a change in interviewer may reflect various non-random processes. Consequently, this significant positive relationship arises due to the combined effect of a whole variety of administration practices, including random changes or changes arising due to

interviewer attrition, interviewer field area changes, household moves, the practice of allocating difficult cases to better performing interviewers and respondent change requests. Therefore, the results for the interviewer continuity variable available in this study – analysing data from FACS, a non-experimental study – should be interpreted with caution.

The recent study by Lynn et al. (2013), which uses experimental data providing control for both interviewer continuity and respondent and interviewer age, shows that the effect of interviewer continuity on the probability of refusal varies by the previous interviewer and respondent characteristics. While interviewer continuity actually reduces refusal propensity for respondents less than 60 years of age if the previous interviewer is over 60, an interviewer change from a young to an old interviewer reduces refusal propensity for respondents over 60 years of age.

II.4.1.2. Interviewer Demographic Characteristics

No significant main effects are anticipated for the demographic characteristics of the interviewer. Although some studies find significant effects for demographic characteristics of interviewers as predictors of nonresponse, the results across studies are not consistent, and other studies including these same variables do not corroborate these relationships. For example, Blom et al. (2010) and Hox and De Leeuw (2002) present significant linear age effects, with higher response rates for older interviewers. On the other hand, the significant categorised age variable in the model predicting individual-level refusal in the study by Campanelli and O'Muircheartaigh (1999) suggests a quadratic relationship between age and individual-level refusal. These different results may be reflecting true differences, or may alternatively be explained in terms of the different interviewer age profiles across the two studies. Whether these significant terms are truly representing age effects or simply vague experience measures – as older interviewers tend to have more years of interviewing experience – cannot be determined with certainty. As Groves and Couper (1998) explain, these differences across studies may arise due to interactions between interviewer and respondent characteristics, a possibility

which will be discussed below, or due to variations across countries and survey topics. These effects may also simply be reflecting differences in experience, personality, skills and behaviour between interviewers having different background characteristics. For example, the higher refusal rates for women found in the study by Hox and De Leeuw (2002) may simply reflect more assertive initial approaches by men or longer years of experience for men due to previously male-dominated interviewing staff. The study by Sinibaldi et al. (2009) supports the idea that gender differences in response rates represent differences in other interviewer characteristics which are differential by interviewer gender. In fact, the authors find a significant bivariate association between interviewer gender and survey cooperation which is explained away by including experience and personality traits variables in the model.

II.4.1.3. Interviewer Experience

A negative monotonic relationship between experience and refusal rates is expected. To the extent that the grade promotions do not occur at the same time for all interviewers, the results of the effect of the 2008 grade on the 2006 wave 8 refusals may indicate some inconsistencies, although the anticipated general trend of lower refusal rates for higher grade interviewers should hold.

The positive influence of the interviewer's experience on the probability of a household to respond has been confirmed across various studies (Hansen, 2006; Hox & de Leeuw, 2002; Pickery & Loosveldt, 2002). Other variants of experience, such as duration in employment on the specific survey considered (Hansen, 2006), years working with current survey agency (Campanelli et al., 1997), and pay grade with current survey agency (Durrant et al., 2010) are also considered in the literature, showing the same relationship as that outlined for total experience. With experience interviewers gain greater confidence in their abilities, and acquire and improve relevant skills, particularly the ability to identify cues about the respondents from their physical appearance, their neighbourhood and their initial reaction and response to the survey request, and the ability to tailor one's approach and behaviour according to these

respondent cues (Groves et al., 1992). Alternatively, a selection effect, with better interviewers remaining in the industry and being promoted, may explain the positive influence of interviewer experience.

One deviation from this general linear relationship finding comes from the study by Durrant et al. (2010), where while a linear positive relationship with cooperation is observed in a model including only interviewer experience, statistical control for interviewer grade changes this relationship to a curvilinear one, with performance dropping slightly for interviewers with very long job experience. This study highlights the importance of considering not only years of experience but also a measure of the position held within the company. However, including both variables may give rise to a collinearity problem. To address these issues, in this study, an interaction variable distinguishing between different experience groups for interviewers holding the same company position is included in the model.

II.4.1.4. Interviewer Work History

The hypothesis for the relationship between work history variables and respondent-level refusal rates is generally that interviewers with greater experience in interviewing and who are in jobs that require skills which are similar to those necessary for interviewing perform better. One would expect that persons who only do interviewing work perform better on the job as they are solely dedicated to this job and consequently have greater practice and are more likely to commit towards the survey agency response targets. The possible relationship between time availability and the probability of response is not clear at this stage and the data analysis will indicate any association. Not all items assessing the interviewers' work priorities, if any, are expected to show any relationship with the outcome variable. The relationship between job satisfaction and performance is hypothesised to be positive, although the size effect is expected to be weak. These variables have not been considered in previous studies, and therefore no evidence of the hypothesised relationships exists yet.

II.4.1.5. Interviewer Personality Traits

There may be no significant relationship between the propensity to cooperate and the allocated interviewer's personality traits. However, if there is any personality effect, the following relationships are conceptualised. It is expected that interviewers scoring high on neuroticism tend to react negatively to stressful situations at the doorstep and this emotional instability may render them ineffective in coping with resistance and persuading reluctant respondents. Extrovert interviewers are expected to engage in communication with sampled respondents with more energy, enthusiasm and confidence, and to react positively to reluctance as a social challenge worth pursuing. Interviewers scoring high on conscientiousness would be expected to be more hardworking, committed to the external benchmarks of survey quality and persistent in achieving the desired cooperation. The impact of being agreeable on an interviewer's performance is not very clear. While respondents may be more likely to cooperate with interviewers who appear honest, trustworthy and sociable, and show understanding of the inconvenience posed, on the other hand agreeable interviewers may be less likely to pressure anyone into participating because they are more likely to consider other people's interests and compromise. It is hard to conceptualise any relationship between openness and interviewer performance.

No consistent pattern has yet emerged from the limited available research on the relationship between the interviewer's personality traits and the propensity of the contacted household to refuse. This may either indicate that the measurement of personality traits may be fraught with error, or that the personality assessment tools used for the general population are not adequate for the analysis of interviewers, or simply that fixed personality traits are not predictive of doorstep interaction success. In this respect, Groves and Couper (1998) hypothesise that tailoring – the most important determinant of success – may be an acquired skill which is independent of fixed personality traits.

Sinibaldi et al. (2009), who utilise the same interviewer survey data as that used in this study, find higher odds of cooperation for extrovert interviewers, and lower odds for agreeable interviewers and interviewers open

to new experiences, significant only at the 10% level, in a multilevel logistic regression controlling for experience, skill and attitudes towards persuasion. The authors justify the higher odds for less agreeable interviewers in terms of their greater reluctance to accept refusals. They also claim that their findings are congruent with those of Snijkers et al. (1999), which indicate higher interviewer response rates for interviewers who assess themselves as not being particularly respondent-oriented and more focused on their task and aims. However, the analysis in the Snijkers et al. (1999) paper simply indicates whether response rates vary by the importance interviewers allocate to different tactics in achieving cooperation. On the other hand, while acknowledging the unexpected direction of the relationship between interviewer openness and respondent cooperation, Sinibaldi et al. (2009) did not put forward any possible explanation for this observed effect. Consequently, the validity of such results is far from conclusive.

II.4.1.6. Interviewer Attitudes

In line with previous studies, it is expected that interviewers who are comfortable with and skilled in using persuasion techniques achieve better response rates. There is evidence in the literature that the interviewers' confidence in their ability to obtain a response even from difficult cases and their attitudes towards the persuasion of reluctant respondents may influence the likelihood of a sampled person agreeing to the survey request. Earlier studies mainly indicate the relationship between such attitudes and self-assurance and interviewer-level refusal rates (De Leeuw et al., 1998; Groves & Couper, 1998; Hox & De Leeuw, 2002). More recent studies consider these interviewer-level variables in the direct analysis of respondent cooperation. Durrant et al. (2010) find that households allocated to interviewers who assert that they are confident in their ability to obtain cooperation, who believe they can achieve a successful result with cases other interviewers find difficulty with, and who disagree with the idea that some people can never be convinced to participate are less likely to refuse. Similarly, Blom et al. (2010) show that households assigned to interviewers who have a positive attitude towards respondent persuasion have higher odds of cooperation. Despite clear

evidence supporting the relationship between response and interviewer confidence and attitudes on persuasion across various studies, Sinibaldi et al. (2009), who used data from the same NatCen interviewer survey used in this analysis, find little evidence of the impact of interviewer attitudes on the persuasion of reluctant respondents on respondent refusal. While significant bivariate associations were confirmed for four out of eight persuasion items in the hypothesised direction, in the multivariate model controlling for weight variables and respondent, survey and area characteristics, only the item *'all can be persuaded'* remains significant.

II.4.1.7. Interviewer Skills

Learnable skills are expected to have a positive effect on cooperation, if any. There seems to be a discrepancy in the evidence of the influence of interviewer behaviour and skills on nonresponse, with studies analysing behaviour directly in the field or the interviewers' perceptions on the importance of different approaches and skills showing significant effects (Morton-Williams, 1993; Snijkers et al., 1999), while studies analysing interviewer self-ratings on fixed operationalised tools do so to a lesser extent (Hox & De Leeuw, 2002; Sinibaldi et al., 2009). This may indicate problems with biased self-assessments or very poor measurement tools which distort results. Sinibaldi et al. (2009) only find one significant skills set – assertiveness – at the 5% level in the multilevel logistic regression predicting cooperation controlling for experience, skill and attitudes. The results indicate a lower probability of cooperation for more assertive interviewers, which is not the expected direction of the relationship.

II.4.2. Model Specification for the FACS Example Dataset

The methods proposed for the analysis of interviewer effects in a longitudinal survey are applied to the FACS data. Due to the changing nature of the sample across waves, and the high number of missing data for previous waves, reflecting administrative failures in the registration of case allocation to interviewers, accurate and complete data for the FACS is only available for the

last two waves. Therefore, the focus is on the last two waves, and models accounting only for both the interviewer from the current wave – wave 8 – and the interviewer from the previous wave – wave 7 – are considered. The respondent area identifier for the FACS dataset is specified at wave 7. The area effect in this context is considered mainly to be the aggregate effect of unmeasured socio-economic and cultural determinants of nonresponse across communities having similar backgrounds. Consequently, a household move between waves should not bring an immediate change in the 'area' effect for that household.

The two most comprehensive models considered for this dataset are a cross-classified model with three distinct random effects – one for area, one for the wave 7 interviewer and one for the wave 8 interviewer – and a MMMC model specifying an area random effect cross-classified with the interviewer multiple membership:

$$\log\left(\frac{\pi_{i(j_8 j_7 s)}}{1-\pi_{i(j_8 j_7 s)}}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_{i(j_8 j_7 s)} + u_{j_8}^{8} + u_{j_7}^{7} + v_s \qquad \text{(II.4)} \qquad \text{and}$$

$$\log\left(\frac{\pi_{i(j_8 j_7 s)}}{1-\pi_{i(j_8 j_7 s)}}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_{i(j_8 j_7 s)} + \mathrm{w}_{ij_8}u_{j_8} + \mathrm{w}_{ij_7}u_{j_7} + v_s, \qquad \mathrm{w}_{ij_8} + \mathrm{w}_{ij_7} = 1 \qquad \text{(II.5)}.$$

In the MMMC model, the overall interviewer random effect is a weighted average effect of the two interviewers that each individual is allocated at wave 8 and wave 7. These weights are represented by w_{ij_8} and w_{ij_7}. While cases allocated to the same interviewer across both waves are given a combined weight of 1 for w_{ij_8} and a weight of 0 for $\mathrm{w}_{\mathrm{i}j_7}$, cases experiencing an interviewer change have two non-zero weights (identical for all cases) summing to 1. Unless otherwise stated, all models are estimated using MCMC estimation in MLwiN Version 2.20 with default priors, a burn-in length of 5,000 and 500,000 iterations. Different burn-in lengths were attempted to identify the appropriate length of discarded iterations to avoid undue influence from the starting values (Gelman et al., 2004). The Brooks-Draper and Raftery-Lewis diagnostics (Browne, 2012) were checked to determine how long the chain must be run for accurate point estimates and 95% credible intervals.

Once an appropriate random structure specification is identified, groups of explanatory variables are considered for inclusion in the following order: participation history, household, area and interviewer variables. Potential predictors are chosen on theoretical grounds and a review of significant predictors in the literature. The order in which the predictor variables for each group are added is determined by the significance of each variable in a separate model (which included the interviewer wave 8 random effect and the between-waves interviewer change variable) using the Wald test. Once all variables pertaining to a particular group are included, their significance is assessed again and variables with a *p*-value less than 0.05 are retained at subsequent steps when including variables from other groups irrespective of their *p*-value thereafter. For all variable groups, excluding interviewer variables, the decision of which variables to include in the model is made on the full dataset (cases with unit nonresponse for the interviewer survey are included). These variables are then forced in the restricted dataset irrespective of whether they become non-significant. When there is a discrepancy in significance at different stages, descriptive statistics for the variables in question have been run. Similar frequency distributions are obtained in the full and restricted dataset, suggesting that non-significance results simply from reduced power. In this analysis the interviewer sex and a grade/experience variable, predictive of nonresponse in the interviewer survey, are included as control variables in the models for the restricted dataset irrespective of their significance.

II.4.3. Exploration of Different Random Effects Specifications

First, simple multilevel models – including only one random effect at a time – are explored. From Table 1 it can be seen that all models indicate significant results for the higher-level random effects. The model including the wave 8 interviewer random effect (Model 3) has the smallest DIC, indicating the best fit, followed closely by the model with the wave 7 interviewer random effect (Model 2), and lastly by the model with the area random effect (Model 1). Higher DIC models correspond to lower effect variance estimates. An important

observation is that the area variance is around half both interviewer variances. All three simple hierarchical models show an improvement on the model with no random effect (Model 0).

Table II.1: Variance and DIC for the Two-level Models

Model	Random Term in the Model	Variance (S.E.)	DIC
0	None		4197.09
1	PSU	0.122(0.051)**	4178.81
2	Interviewer 7	0.233(0.065)**	4155.98
3	Interviewer 8	0.273(0.077)**	4150.64

** significant at the 1% level

Multilevel cross-classified models including only one of the interviewer random effects in addition to the area effect are considered next (Table 2, Models 4 and 5). For both models, the area random variance is no longer significant, and the variance is reduced to about 40% of the variance estimated for the 2-level model including only the area effect (Model 4 and Model 5 compared to Model 1). Interestingly, the interviewer variance estimated in a model cross-classified with area is about 0.9 times the random effect obtained in a model controlling only for the interviewer random effect (Model 4 compared to Model 2; Model 5 compared to Model 3). This reduction suggests that although there may be some confounding if the area effect is eliminated, this is minimal. In fact, the DIC statistics for Models 4 and 5 are only slightly lower than those obtained for the equivalent Models 2 and 3 controlling only for the respective interviewer effect, indicating that these cross-classified models do not offer a noticeable improvement to the simpler 2-level models.

A cross-classified model controlling for both interviewers at wave 7 and 8 but excluding area effects (Table 2, Model 6) yields numerically unstable results. Stability is reached only once the MCMC chain length is increased to 5 million iterations, at which point both effects are only just about significant. The variance for each interviewer random effect is reduced to around 0.6 of the variance observed in the two separate models including only one random effect (Model 6 compared to Model 2; Model 6 compared to Model 3). For these

models, the wave 8 interviewer random effect always has a slightly lower *p*-value than that for interviewer 7. Model 6 has a DIC value that is four points lower than the model including only the wave 8 interviewer random effect (Model 3). The cross-classified model with all three random effects (Table2, Model 7) did not converge. The high negative correlation between the two variance estimates can be observed visually in the diagnostics trajectories graphs. The chains show very poor mixing. The effective sample sizes obtained for the area, current interviewer and previous interviewer variance parameters are 754, 502 and 344 for runs of 500,000 iterations, indicating poor mixing of the chain. The second order Penalised Quasi-Likelihood estimate could not be estimated either as the variance and covariances matrix turned negative definite. Despite attempts to approximate this matrix to the nearest positive definite matrix convergence could not be achieved, possibly indicating that the model was incorrectly specified, or that a parameter needed to be excluded. The same problem is encountered when using first order Marginal Quasi-Likelihood (MQL) estimates as starting values.

Table II.2: Variance and DIC for the Cross-classified Models

Model	Random Terms in the Model	Variance (S.E.)	DIC
4	Interviewer 7 & PSU CC	0.210(0.068)**; 0.048(0.044)	4153.77
5	Interviewer 8 & PSU CC	0.247(0.079)**; 0.047(0.044)	4149.79
6	Interviewer 7 & Interviewer 8 CC	0.139(0.080)*; 0.167(0.095)*	4146.27
7	Interviewer 7, Interviewer 8 & PSU CC	did not converge	

* significant at the 5% level; ** significant at the 1% level
CC Cross-classification

The instability and substantial reduction in estimates in Model 6 suggest that the two interviewer random effects are near non-identifiable. As anticipated, this may indicate that the assumption of independent interviewer effects is erroneous, resulting in a misspecification of the interviewer-level structure. The difference in the random effect estimates for the two wave

interviewers may be simply reflecting interviewer changes between the two waves, rather than a differential effect of the previous wave interviewer from the effect of the current wave interviewer on the propensity to refuse in the current wave. In fact, when a fixed effect for the interviewer change variable is included in the cross-classified model with the two interviewer random effects, one of the effects becomes non-significant.

Alternatively, multiple membership models are explored using a range of weights (Table 3). As described in the methodology section, cases being allocated to the same interviewer across both waves were given a weight of 1 for the wave 8 interviewer and a weight of 0 for the wave 7 interviewer. The weights allocated for cases experiencing an interviewer change are specified in Table 3. The results show that the multiple membership models which give a weight of at least 0.8 to wave 8 interviewers provide a better model (Models 15 & 16). These results support the hypothesis that the wave 8 interviewer has the greatest impact on the current wave nonresponse. The next best fitting models are obtained for a high wave 7 weight (Models 13 & 14); these perform better than models allocating moderate weights (approaching an equal share) to both wave interviewers (Models 8, 9 & 10). This would seem to suggest that there is not much difference between the wave 7 and wave 8 interviewer effects.

The MMMC model (Model 17) with the preferred weight specification – 0.9 and 0.1 for wave 8 and wave 7 respectively – shows only a slight improvement in the model fit compared to the multiple membership model that does not account for the area cross-classification (Model 16). However, the area random effect in this model is not significant. Comparing the DIC values of Models 16 and 17 to the value of Model 3 suggests that for our application, the multiple membership models do not provide a substantial improvement to the simpler 2-level model, only accounting for the current wave interviewer (wave 8).

Table II.3: Weights, Variance and DIC for the Multiple Membership and MMMC Models

Model	Random Terms in the Model	Wave8, Wave7 Weights	Variance (S. E.)	DIC
8	MIM	0.4, 0.6	0.278(0.086)**	4159.08
9	MIM	0.5, 0.5	0.287(0.087)**	4159.03
10	MIM	0.6, 0.4	0.288(0.090)**	4158.75
11	MIM	0.3, 0.7	0.272(0.082)**	4158.33
12	MIM	0.7, 0.3	0.291(0.090)**	4157.41
13	MIM	0.2, 0.8	0.262(0.076)**	4157.38
14	MIM	0.1, 0.9	0.250(0.074)**	4155.85
15	MIM	0.8, 0.2	0.288(0.085)**	4155.36
16	MIM	0.9, 0.1	0.282(0.081)**	4153.12
17	MIM & PSU CC	0.9, 0.1	0.252(0.084)**; 0.049(0.049)	4151.62

** significant at the 1% level

MIM multiple interviewer membership; CC Cross-classification

In conclusion, for this application, simply controlling for one wave interviewer may be sufficient. The area random effect is shown across all models to be negligible once the interviewer random effect is controlled for, warranting its exclusion from the model. The 2-level models indicate that the wave 8 interviewer random effect (Model 3) produces a slightly better model fit than the model with wave 7 interviewer effect (Model 2), and also that it explains a larger proportion of the total variance. Consequently, the final model only accounts for the wave 8 interviewer clustering.

II.4.4. Discussion of the Final Model – Random Effects

Table 4 presents estimates of the interviewer wave 8 random effect as groups of explanatory variables are added to Model 3, for the final dataset of 5932 cases. Also included are the percentage reduction in variance compared to the initial variance obtained in the null model (a) or compared to the variance obtained in the preceding model (b), the DIC and the VPC values. The DIC values indicate a better model fit at each step of the model building exercise, as more significant fixed effects are included in the model.

Table II.4: Estimates of the Interviewer 8 Random Effect Variance as Groups of Explanatory Variables Are Added

Fixed Effects Parameters (no. of parameters)	Variance (S.E.)	Percentage Reduction (a)	Percentage Reduction (b)	VPC	Model DIC
None	0.279(0.088)**	n/a	n/a	7.8	3397.7
Added Interviewer Change (1)	0.198(0.081)**	29.0	29.0	5.7	3385.6
Added Participation History (2)	0.179(0.078)*	6.8	9.6	5.2	3363.6
Added Respondent/Household Characteristics (12)	0.158(0.083)*	7.5	11.7	4.6	3336.6
Added Interviewer 8 Sex & Grade/Experience (7)	0.105(0.076)	19.0	33.50	3.1	3327.2
Added Interviewer 8 Work History (4)	0.090(0.070)	5.4	14.3	2.7	3316.4
Added Interviewer 8 Personality Trait (3)	0.057(0.057)	11.8	36.7	1.7	3309.8
Added Interviewer 8 Skills (4)	0.044(0.049)	4.7	22.8	1.3	3303.0

Percentage reduction a - percentage reduction in variance compared to the initial variance 0.279(0.88)

Percentage reduction b – percentage reduction in variance compared to the variance obtained in the preceding model

The interviewer random effect remains significant until the inclusion of the interviewer administrative variables: sex and interviewing grade/experience. The variables included in the model explain all of the interviewer random effect. This supports the findings by Campanelli and O'Muircheartaigh (1999) who find interviewer effects to be no longer significant once fixed effects are controlled for. For the first model, not including any fixed effects, the interviewer variance accounts for around 7.8% [0.279/(0.279 + 3.29)] of the total variation in refusal at wave 8. Once variables for interviewer change, participation history, household-level, interviewer experience and grade, work history, personality traits and skills are controlled for, the VPC is reduced to 1.3%. Although the interviewer variance also decreases slightly when including the twelve household-level fixed effects, the more substantial decreases come from the interviewer-level variables. The reduction in interviewer variance with the inclusion of household characteristics can be explained in terms of

systematic differences in the allocation of cases to interviewers. Also, since the individual-level variance is not allowed to vary, there may be a scaling effect when introducing new variables in the model. When individual-level variables are added to a multilevel logistic model the level-one variation cannot decrease since the residual variance at the individual level is fixed. Instead the estimates of the other regression coefficients and the intercept correlation will change by a specific factor (Snijders & Bosker, 1999). The interviewer change variable and the interviewer grade and experience variable are responsible for the greatest reduction in the interviewer random effect. Participation history also reduces the interviewer variance substantially, which may reflect different work allocations across waves by interviewer.

Before deciding on the final specification of the multilevel model, multiple membership models with different weight specifications including all relevant explanatory variables are revisited to evaluate their performance in comparison to the simpler 2-level model (Table 5). However, no noticeable differences in the DIC values across the various models are found, implying that, for this application, a simpler 2-level model is indeed sufficient, even after the inclusion of explanatory variables. An attempted swap between the interviewer random effect with an area random effect, while maintaining all fixed effects, results in a higher DIC value and a non-significant area effect.

Table II.5: Random Effect Estimates for a Two-level and Multiple Memberships Random Effect Specifications for the Final Choice of Fixed Effects

Random effect	W8, W7 Weights	Variance (S.E.)	VPC	Model DIC
Wave 8 interviewer		0.044(0.049)	1.3	3303.00
Interviewers multiple membership	0.9, 0.1	0.048(0.050)	1.4	3302.94
Interviewers multiple membership	0.8, 0.2	0.046(0.052)	1.4	3303.61
Interviewers multiple membership	0.7, 0.3	0.041(0.049)	1.2	3303.49

II.4.5. Discussion of the Final Model – Fixed Effects

Table 6 presents the estimated coefficients of the final multilevel logistic regression model. The non-significant random effect was retained in the final model. The removal of this non-significant random effect does not vary the estimates and the interpretation of results substantially. Below the significant variables in the model will be discussed with reference to previous findings and with a clear focus on the possible implications for practice. Alternative transformations of the variables included in the final model, their results and interpretations, and the reasons for discarding these alternatives are also discussed.

II.4.5.1. Significant Participation History and Household-Level Variables

Participation history variables – a variable indicating the first wave the case was included in the sample and a variable specifying the number of times the case was interviewed – are highly significant when included one at a time in the model. Both variables cannot be retained in the model simultaneously because of their large correlation with each other. Consequently, the variable obtaining the lower *p*-value for the Wald test in a multilevel model including only the interviewer wave 8 random effect and the interviewer change fixed effect is retained. Therefore, the variable 'First Wave' is included in the final model.

In comparison to cases introduced into the sample in wave 7, cases introduced to the survey at an earlier wave show a lower propensity to refuse the survey request. This result is significant for all prior six wave dummy variables at the 5% level. This indicates that participants who have been in the sample for a greater number of waves are generally more likely to continue participating at later waves. This suggests that at successive waves, participants who remain in the survey constitute a selective population that is committed to the survey request and is more likely to maintain its commitment throughout. Similar results were obtained by Nicoletti and Peracchi (2005) who show that the probabilities of response conditional on contact for the second,

third and fourth interviews are higher than that for the first interview, and that this positive influence increases for every additional interview.

The dummy variables for Wave 1, Wave 2, Wave 3 and Wave 4, as well as the dummy variables for Wave 5 and Wave 6, have similar coefficients which are not significantly different at the 10% level. Consequently, these two groups have been collapsed into two dummy variables: Waves 1-4 and Waves 5-6. This variable transformation still confirms the previous results of a lower likelihood of refusals for respondents who were included in the sample at an earlier wave. The results indicate a significant increase in the likelihood of cooperating in the next survey contact from those having participated in the survey once or twice and those who participated at least three times before, at the 1% level. While participants introduced to the survey at Waves 1-4 have 0.41 times the odds of refusing participation at wave 8 when compared to those introduced at Wave 7, for those initiated at Waves 5-6 the odds are 0.64 times smaller.

The variable discussed above does not indicate the real first wave for all cases. Cases which have been re-entered at a specific wave have only been coded at that wave, and therefore, for such cases the true first wave is not available. However, this amounts to only 160 cases of the full dataset. Due to these cases, a variable indicating the proportion of waves in which the respondent cooperated – obtained by dividing the number of successful interviews out of the number of waves the case was included in the sample, and which could have provided a way of including information from both participation history variables – has not been included. Consequently, the results obtained in this study cannot be compared with those obtained by Campanelli and O'Muircheartaigh (1999), Watson and Wooden (2006), and Laurie et al. (1999), which show that partial nonresponse in the previous wave is associated with higher refusal propensity at a later wave at both the household and individual level.

As explained previously, the main focus of the paper is on interviewer effects on nonresponse. Consequently, the results for respondent- and household-level predictors are reviewed only briefly below.

Table II.6: Estimated Coefficients for the Final Multilevel Logistic Model Analysing Wave 8 Nonresponse

Variable {Reference Category}	Category	B	S.E.	β/S.E.	p-value
Interviewer Change {Same}	Change	0.409	0.110	3.718	0.000
	Participation History Variables				
First Wave for Respondent {Wave 7}	Wave 1-Wave 4	-0.892	0.145	-6.152	0.000
	Wave 5-Wave 6	-0.443	0.163	-2.718	
	Respondent/Household Variables				
Ethnicity {Non-white & Missing}	White	-0.524	0.155	-3.381	0.001
Any Vocational or Academic Qualifications {Yes}	No	0.289	0.144	2.007	0.044
Age of Youngest Child {No dependent children & 16-18 year olds}	0-4 years	-0.548	0.170	-3.224	0.004
	5-10 years	-0.441	0.171	-2.579	
	11-15 years	-0.189	0.175	-1.080	
Heating Problems in the Dwelling {No & Don't know}	Yes	0.266	0.217	1.226	0.225
Gender {Female}	Male	-1.428	0.554	-2.578	0.010
Accommodation Type {Detached house}	Semi-detached house	-0.280	0.125	-2.240	0.198
	Terraced house	-0.264	0.137	-1.927	
	Flat or maisonette - purpose built & Other	-0.159	0.211	-0.754	
	Flat or maisonette - conversion	-0.504	0.523	-0.964	
Household Size	Household size	0.086	0.046	1.870	0.056
	Interviewer 8 Administrative Variables				
Gender {Female}	Male	0.094	0.111	0.847	0.391
Grade Experience {Grade R; S, 5+ years' experience; T (Highest grade)}	Grade A (Lowest Grade); B	1.117	0.278	4.018	0.000
	Grade C, 0-4 years' experience	1.027	0.212	4.844	
	Grade C, 5+ years' experience	0.495	1.640	0.234	
	Grade D, 0-4 years' experience	0.812	2.252	0.254	
	Grade D, 5+ years' experience	0.530	1.699	0.252	
	Grade S, 0-4 years' experience	0.701	2.016	0.286	

Interviewer 8 Work History, Time Availability, Attitudes towards Refusal, Work Priorities, Satisfaction with Job Variables

Interviewing Work History – Work status with another survey agency & Experience with other (phone, marketing) survey interviewing {Never worked for another survey agency}	Working with another agency at time of survey & Done other survey interviewing	-0.208	0.174	-1.195	0.000
	Worked with another agency prior to 01/01/06 & Done other survey interviewing	0.436	0.131	3.328	
	Working with another agency at time of survey & Never done other survey interviewing	-0.610	0.296	-2.061	
	Worked with another agency prior to 01/01/06 & Never done other survey interviewing	0.241	0.185	1.303	
	Interviewer 8 Personality Traits Variables				
Worries a lot 1 {Does not apply to me at all)}	2, 3	0.582	0.204	2.853	0.002
	4	0.298	0.220	1.355	
	5, 6, 7 (Applies perfectly to me)	0.690	0.213	3.239	
	Interviewer 8 Skills Traits Variables				
Expresses myself easily {1 (Does not apply to me at all), 2, 3}	4	0.342	0.291	1.175	0.011
	5, 6, 7 (Applies perfectly to me)	0.638	0.251	2.542	
Can't help but look upset when something bad happens {6, 7 (Applies perfectly to me)}	1 (Does not apply to me at all), 2	-0.28	0.17	-1.709	0.053
	3, 4, 5	-0.33	0.14	-2.416	

The odds of non-white sampled persons refusing participation are 1.7 times those of white sampled persons, and this result is significant at the 1% level in the final model. Any differences between the various non-white ethnicities are not significant, and therefore, these categories have been collapsed together. This result is congruent with the analysis carried out by Campanelli, Sturgis and Purdon (1997), who use the same categorisation for ethnicity. The odds of male respondents refusing to participate are roughly 75% lower than the odds of a female respondent. No conclusive evidence of the

effect of gender has been found in the literature. Some studies in the literature found no relationship between refusal and sex (Watson & Wooden, 2006), some found higher cooperation (Hox & De Leeuw, 2002), while others found lower cooperation rates for women (Pickery & Loosveldt, 2002). Therefore, the relationship between sex and refusal varies by survey topic and country. In this case, since the study is mainly targeted at the mother figure, single fathers or male guardians may have a special interest in participating in the survey and voicing their views.

The odds of respondents with no vocational or academic qualification refusing are 1.4 times the odds of respondents with some educational attainments. Further disaggregation of the education variable does not indicate variations in the probability of refusal between respondents with different levels of educational attainment. The literature on the relationship between the respondent's education level and the likelihood of cooperation is mixed. For example, while Groves and Couper (1998) find a negative relationship between education and cooperation described, Watson and Wooden (2006) find a significant relationship in the opposite direction.

The final model indicates higher rates of refusals for larger households for the UK FACS. However, this result is only significant at the 10% level in the final model. This deviates from the study by Haunberger (2010) which included the same variable in two separate models analysing refusal at wave 2 and wave 3 for the German Child Longitudinal Study, but did not find any significant effect. The response analysed in Haunberger (2010) is also that for the mother figure within the household. While in a general household survey, where any adult within the household can participate in the survey, the likelihood of response may be expected to be higher, for a survey where the mother figure (or father figure where the mother figure is missing) is the only eligible respondent a bigger household may translate into a lower probability of response due to a greater workload burden. Consequently, the variations in results may be explained in terms of contextual differences between Germany and the United Kingdom, specifically in relation to the division of household duties and childcare arrangements. A categorical transformation of this continuous variable was attempted, owing to evidence of non-linear effects

cited in the literature (Watson, 2003). The results suggest that a linear relationship (on the logistic scale) between household size and refusal exists, with increasing refusal probabilities for categories showing larger household sizes. Since the continuous variable obtained a better model fit than its categorical equivalent, the continuous measure has been retained in the final model. On closer inspection, the non-linear effects found by Watson (2003) show higher attrition rates for one-person and larger households. In fact, the higher refusal rates for single person households have been widely documented. However, the dataset used in this study does not include households with only one person, because of the focus of the study on households with children, eliminating the lower end of the distribution, and the quadratic relationship.

Similarly to Watson (2003), households with dependent children are more likely to cooperate with the survey request, while the number of dependent children does not significantly predict refusal. Similarly, Nicoletti and Peracchi (2005) find that the number of children in the household is not associated with the probability of cooperation given contact, while they confirm a positive relationship between this variable and contact. On the other hand, the variable indicating the age of the youngest child shows that the odds of refusal are lowest for households with younger children, possibly due to the fact that the eligible respondent may spend more time at home due to caring duties when younger children are present, and consequently would be more available for an interviewer. Households with 11 to 15-year-old children, despite showing lower odds of refusal, are not significantly different in their refusal probabilities from households with no dependent children or children aged sixteen years, whilst showing significantly higher odds at the 1% level compared to households with 0 to 4-year-olds, and significantly higher odds at the 10% level compared to households with 5 to 10-year-olds.

Some variables included in the final model are not significant, due to the modelling strategy chosen, as explained in the methodology section. The identification of heating problems acts as a proxy for household deprivation. Such acknowledged deprivation seems to increase refusals; however, this is not significant in the final model at the 10% level. The variable accommodation

type as a whole is also no longer significant in the final model. However, there are significant differences between respondents living in semi-detached and terraced houses compared to those living in detached houses, with respondents living in detached houses showing higher refusal rates.

II.4.5.2. Significant Interviewer Fixed Effects

The following discussion focuses on the effects of interviewer-level variables on nonresponse.

II.4.5.2.1. Interviewer Continuity

In agreement with previous observational studies (Schatteman, 2000; Watson & Wooden, 2006; Haunberger, 2010), an interviewer change between waves is positively associated with refusal. However, a causal relationship cannot be inferred as an interviewer change may reflect a respondent move or the resignation of an interviewer rather than a random allocation (Hill & Willis, 2001).

With the aim of identifying whether the effect of a change of interviewer depends on the grade, a more complex categorisation for change has been included for exploratory purposes in a model controlling only for the wave 8 interviewer random effect. This variable distinguishes between cases experiencing change where the wave 8 interviewer is of the same, higher and lower grade group as the wave 7 interviewer. The grade groups are A, BCD, RT, and S, with A being the lowest grade and T the highest. The estimates of these dummy variables seem to indicate that there are no significant differences between cases with no change and changes for a higher grade group interviewer, while cases experiencing a change to an interviewer of the same or lower grade group are more likely to refuse at the 1% significance level. This result hints to the possibility of improving the odds of cooperation at the subsequent wave by purposely changing the interviewer to a substantially more highly experienced one.

Additionally, another variant for the interviewer continuity variable has been constructed. This variable distinguishes between cases experiencing an interviewer change specifically due to the wave 7 interviewers completely dropping out of the survey at wave 8 and all other interviewer changes. In a simple hierarchical model including only the wave 8 interviewer random effect, the dummy variable for changes arising from interviewer attrition is significant only at the 10% level, indicating higher refusal rates for cases allocated to a new interviewer at wave 8 owing to the attrition of their previous interviewer in comparison to cases with no interviewer changes. However, once the grade and experience variable for the wave 8 interviewer is included in the model, this dummy variable is no longer significant. This would seem to indicate that interviewer changes arising from interviewer attrition do not influence the sampled person's propensity to cooperate. This result suggests that the negative effect of interviewer change on cooperation identified in the final model may be due to cases where the respondent has moved house and is therefore more likely to consider the survey to be an inconvenience during a particularly stressful time for the household, or in situations when the previous interviewer suggests a change in job allocation due to difficulties encountered with the specific case.

II.4.5.2.2. Interviewer Grade and Experience

Both interviewer grade and years of experience are highly significant predictors when included one at a time in the model. While years of experience simply constitute a consistent measure of exposure in the profession, the grade held within the organisation is a more subjective construct and may reflect experience, educational background and skill. In fact, while interviewers with higher grades are generally more likely to have many years of experience, for a particular grade there are interviewers with a range of years of experience. However, both variables are positively correlated and their simultaneous inclusion would cause problems with collinearity. Consequently, a variable distinguishing between different years of experience bands for the same grade has been created. The categorisations reflect differences in refusal rates across

different experience and grade groups, as well as necessary collapsing of categories owing to small sample sizes.

All categories of the grade/experience variable show a significantly higher propensity of refusal compared to the highest grade/experience group at the 5% level. As hypothesized, interviewers in the lower experience group for both grades C and D have higher odds of refusal than those interviewers in the higher experience group for the same grade. The difference between these groups is, however, only significant for grade C. The positive effect of grade and experience is a consistent effect confirmed across various studies (Campanelli et al., 1997; Hox & de Leeuw, 2002; Pickery & Loosveldt, 2002; Hansen, 2006; Durrant et al., 2010), and may either indicate improved performance over time and as one moves up in the company hierarchy, or a selection effect with better interviewers remaining in the industry and being promoted.

II.4.5.2.3. Interviewer Work History

In the final model, the work history variable shows that terminated work experience with another survey agency and experience in other types of interviewing have a negative effect on individual-level cooperation. The negative effect of work experience in different interviewing modes or research areas suggests that face-to-face interviewing requires specific tactics and skills, and that exposure to techniques suitable for other types of interviewing may hamper performance. The results indicate that interviewers working with another survey agency at the time of the survey performed better than those who had previous experience with another survey agency and better than those with no such experience. This result may indicate that interviewers who commit most of their paid working hours undertaking interviewing work for various survey agencies perform best. The explanation for the negative effect of previous work with other survey agencies is somewhat unclear and more data on interviewing work history is required to explore this relationship further. Possibly, job tenure in interviewing work shows commitment and skill

in one's work, and consequently interviewers who are no longer working for other survey agencies may represent less able or less committed interviewers.

Variables indicating whether the interviewer had undertaken any other type of work – including other (phone or marketing) types of interviewing, interviewing for another survey agency, and non-interviewing work – are all significant when included one at a time. Only one of these variables could be retained in the final model because of the high correlation between these variables. Those who have never done other types of interviewing are more likely to have never undertaken interviewing with another survey agency. Contrary to what might be expected, those who have another job (other than interviewing) are also more likely to be working with another survey agency. This could indicate some problems with the interpretation of the question relating to whether interviewers had another non-interviewing job, and consequently, for the purposes of this study, this variable is excluded.

When the other two variables are included simultaneously in the model, the following results are obtained: interviewers who had not undertaken other survey interviewing obtain significantly lower respondent-level refusal rates than interviewers who had; and interviewers who had worked with another survey agency for at least two years, and to a lesser extent (smaller effect size) those who are currently working for another survey organisation at the time of the survey, show significantly lower refusal rates than interviewers who had only ever worked for NatCen. On the other hand, interviewers who had worked for another survey agency prior to 1st January 2006 do not significantly differ from those who had only worked for NatCen.

For these two variables, an interaction variable has been created by summarising the interviewer work history, and this is the variable included in the final model, discussed above. Due to the low number of interviewers for some categories of this interaction variable, the distinction between those currently working for another survey agency at the time of survey and who had started work with this agency at a later date than 1st January 2006, and those who were currently working for another survey agency, having worked there at least since 1st January 2006, is removed.

II.4.5.2.4. Interviewer Personality Traits and Skills

Personality traits and interviewer skills have been hypothesized to play a role in explaining interviewer performance (Weinhardt & Kreuter, 2011). In the final model only individual items, treated as categorical variables, are included. In considering each personality trait item individually, score categories are collapsed together in cases where the coefficient of each category is not significantly different from the other or the number of cases in each category is too low. The requirement for the groupings is easy interpretation and is restricted to neighbouring scores only.

Here, only one personality item is retained in the final model. This item indicates that respondents who are allocated interviewers with a low or high self-rating of neuroticism in terms of worrying tendencies are more likely to refuse participation than those respondents allocated to interviewers who assert that this item does not describe them at all. On the other hand, interviewers showing a moderate score of 4 are not significantly different to the reference category at the 5% level.

Two skills items are retained in the final model – 'can't help but look upset when something bad happens' and 'express myself easily'. Contrary to expectations, respondents approached by interviewers rating themselves as highly capable of expressing themselves with ease (5, 6, 7) are significantly more likely to refuse than those allocated to interviewers who rated themselves poorly (1, 2, 3) on this skill, while there is no significant difference for those interviewers with moderate scores (4). This result possibly indicates that interviewers who are less complacent about their ability to convey the survey message, who have greater awareness of the way they portray themselves, and who make a conscious effort to communicate effectively achieve higher response rates. Sample cases allocated to interviewers who perceive themselves as never allowing others to notice they are upset when something goes wrong are less likely to obtain respondent-level refusals than interviewers who recognise they are very likely to show such feelings. However, this result is only significant at the 10% level. On the other hand, interviewers scoring moderately (3, 4, 5) on this item are significantly less likely to obtain respondent-level refusals compared to the highly transparent group. This

result might highlight that interviewers who do not get flustered or defeatist if a sampled person shows scepticism or hesitancy have more chance of success.

Of substantive interest is the recurrent pattern for personality and skills items, where interviewers providing moderate answers scored generally better than those providing extreme values. This may indicate that the most confident interviewers do not necessarily perform best on the job. It may also be plausible that while interviewers who are confident in their performance may have been more likely to tick moderate scores on the traits items in the interviewer questionnaire, others were more subject to social desirability bias and tended to overrate their personality disposition and skills for the job.

Besides attempting to include each personality trait item individually, both as continuous and categorical variables, an alternative method for including personality traits in the model was attempted. This method is the same used by Sinibaldi et al. (2009), which consists of taking the mean of the three traits pertaining to each personality dimension. When each personality trait is included in a simple hierarchical model controlling only for the wave 8 interviewer random effect and the interviewer change fixed effect, only 'Extroversion' – showing lower refusals for more extrovert interviewers – and 'Openness' – unexpectedly showing higher refusals for more open interviewers – are significant as linear effects. These results are very similar to the effects obtained in the study by Sinibaldi et al. (2009). When the grade/experience variable is controlled for, neither of these personality traits remain significant, achieving *p*-values greater than 0.10. The difference in results for the Sinibaldi et al. (2009) paper and this study may simply be due to the fact that the personality of the interviewer plays a more important role in gaining response in cross-sectional surveys than later waves of a longitudinal survey. As there may be non-linear effects for these personality traits, with both very low and very high scores having the same effect on nonresponse, an attempted alternative specification of these traits is the inclusion of a quadratic term. However, none of these terms are significant.

Similarly to the personality traits items, the inclusion of skills items both individually, as continuous and categorical variables, and as composite

measures was attempted. In this study the 35 of the 52 skills items associated with the ability to achieve response rather than contact are considered (Sinibaldi et al., 2009). Creating a composite measure using a principal component analysis requires the complete interviewer dataset. Then the component scores and loadings obtained in this analysis of all interviewers can be applied to the interviewers included in this study. This strategy reflects the understanding that the underlying dimensions for these traits should be uncovered for the population of interviewers. However, this data is not available and the dimension reduction analysis could either to be carried out on the restricted sample or by using the results obtained by Sinibaldi et al. (2009). The authors use data from 845 interviewers out of a total of 1198 interviewers who participated in the interviewer survey, making their dataset more complete than the one available for this analysis for determining the correlations between different skills items. Consequently, the components and loadings available in their study are applied in this paper. The authors have run a principal component analysis, followed by a confirmatory factor analysis which indicates that the 35 items can be grouped in ten components.

In a 2-level model controlling only for the wave 8 interviewer random effect and the interviewer change fixed effect the factor 'Emotional Resilience' is the only factor reaching significance, showing lower refusal rates for interviewers with higher scores. However, when the variable grade/experience is controlled for this variable is no longer significant, not even at the 10% level.

II.4.6. Discussion of the Non-significant Fixed Effects

II.4.6.1. Area Effects

The variables describing the geographical area of the household, such as the indicator for the UK regions, the London indicator, and various respondent neighbourhood perception variables are found not to be significant after controlling for other household-level variables, confirming similar findings in Durrant et al. (2010). The inclusion of these variables in a model also controlling for household-level variables and participation history does not result in any substantial reduction of the interviewer-level variance. This result

supports the conclusion that, after controlling for household and interviewer effects, area effects are negligible.

II.4.6.2. Interviewer Effects

The working field area variable, indicating the geographical area within which the interviewer may be allocated work, is not significant. This provides further evidence that there are no significant area effects on nonresponse. Demographic variables such as gender and age are also not significant. The evidence of demographic interviewer effects in previous literature is mixed. Variables indicating the importance interviewers allocate to various aspects of a job, such as monetary compensation and flexibility, are all not significant.

The untransformed items indicating the interviewer attitudes on the persuasion of reluctant respondents – as considered in the study by Sinibaldi et al. (2009) – are not predictive of respondent refusal. The effect of attitudes on respondent persuasion may be more important for cross-sectional surveys, investigated in the study by Sinibaldi et al. (2009), when sampled members are being approached to take part for the first time, in comparison to later waves in a longitudinal study, being investigated in this study, where sampled members already know the survey and its scope. In the current study, the inclusion of each persuasion item was attempted individually and then collectively in a 2-level model controlling only for the wave 8 interviewer random effect and the interviewer change fixed effect, but none are significant. In a similar manner to the method used in the study by De Leeuw et al. (1998), an attitude index is constructed to aggregate the scores across the eight different items which indicates greater reluctance to use persuasion techniques and to respect the voluntary nature of the participation request for higher values. This index also is not significant in the simple hierarchical model. These results are contrary to various studies showing that an interviewer's positive attitude towards the effectiveness and acceptability of respondent persuasion is conducive towards achieving higher response rates (De Leeuw et al., 1998; Groves & Couper, 1998; Hox & De Leeuw, 2002). However, the study by Sinibaldi et al. (2009), which is analysing a larger pool of interviewer from

the same survey agency considered in this study, finds only weak evidence of the effect of such attitudes on respondent cooperation: only one item is significant at the 5% level in the final model, indicating higher respondent-level cooperation rates for interviewers who believe that with enough effort all respondents can be convinced to participate.

All variables indicating the interviewer satisfaction rating with various aspects of the survey agency except one are not significant at any stage in the multilevel model. The variable enquiring about the overall satisfaction with the job with the NatCen survey agency is significant at the 10% level in a 2-level model controlling only for the wave 8 interviewer random effect and the interviewer change fixed effect. Respondents allocated to interviewers belonging to the 'Quite Dissatisfied and Very Dissatisfied' category show a significantly higher likelihood of refusal compared to the respondents interviewed by the 'Very Satisfied' interviewer group. However, neither the 'Quite Dissatisfied' nor the 'Neither Satisfied nor Dissatisfied' group has a significant effect, despite both having positive coefficients. The notion that interviewers perform better if they are satisfied with their working conditions in the agency is understandable. On the other hand, it may be that better performing interviewers are more satisfied with their work simply because they are happy with their achievement or because they are given more work and are better respected for their performance. When this variable is included in a model controlling for other significant variables, it is no longer significant.

A variable indicating the workload of each interviewer has been created. When included in a 2-level model controlling only for the wave 8 interviewer random effect and the interviewer change fixed effect this variable shows a significant negative association with nonresponse at the 5% level, indicating lower odds of refusals for cases allocated to interviewers with a greater case burden. Nicoletti and Buck (2004) find workload to have a highly significant negative relationship with cooperation, indicating that huge work burdens negatively affect work quality. Watson and Wooden (2006) attempt both linear and quadratic specifications of workload to account for both the effect confirmed by the above-mentioned study and an opposing effect arising from the management decision specific to their survey of allocating greater

workload to more experienced interviewers. The authors report a significant quadratic effect for the contact propensity model with an optimal workload of 124 cases per interviewer and worse contact outcomes for lower and higher caseloads, but no significant effect is found for response. As both possible contradictory workload effects may be applicable to the current study, the inclusion of a quadratic term has been considered, but this is not significant.

Since the dataset is conditional on wave 7 cooperation, this measure of workload does not account for opt-in cases or booster cases at wave 8. This is the same restriction to the workload measure considered in the study by Watson and Wooden (2006). To the extent that panel, opt-in and booster cases are not distributed in equal proportions across interviewers, particularly interviewers of different experience bands, the interpretation of the association between refusals and interviewer workload may be more complex than apparent. To the extent that new interviewers are mainly allocated more new cases such as boosters or opt-in cases, the interviewer workload will show them as having very low workload, and consequently this effect may simply be masking the lower refusal rates for interviewers holding higher grade positions within the company and with longer years of experience. In fact, interviewer workload is highly correlated with grade and experience, with interviewers in the higher grades and with greater experience generally having a greater number of cases registered for this restricted definition of workload. When controlling for the grade/experience variable, this variable becomes barely significant at the 5% level. To avoid multicollinearity this variable has been excluded from the model.

Several variables indicating the time availability of interviewers have been constructed. Separate indexes are composed for those dummies indicating time slots during which interviewers are otherwise occupied and those during which they are unwilling to do the interviewing. Several variants of the time availability and willingness constructs are produced. One variable simply gives the same weight to any time slot and sums up all those time slots which the interviewer indicates as inconvenient. Another variable gives a double weight to weekend slots. Two other variables only consider specific time slots – one considers weekend (Saturday and Sunday) slots, while another

considers weekend and evening (6-9 pm) slots. For all these variables, missing answers for particular time slots are considered as convenient times, and therefore, not contributing to the index of indisposition for that interviewer. A transformation of these indexes from continuous to categorical scale is also attempted.

The only significant time restraint construct in a 2-level model controlling for the wave 8 interviewer random effect and the interviewer change fixed effect is the unwillingness index allocating equal weight to each time slot and categorising scores in 3 groups: 0, 1-10, 11+. This variable suggests that while interviewers who are unwilling to work during up to ten time slots obtain worse cooperation rates, very inflexible interviewers with very limited working hours perform better. Possibly interviewers who, despite being unavailable during most of the working times, are still allowed to work for the agency are experienced interviewers who have other work commitments but are retained within the organisation for their excellent cooperation rates obtained. A cross-tabulation of the index score categories by interviewer grade and interviewer experience separately confirms the tendency for higher grade and higher experienced interviewers to be less available for interviewing work. The variable is no longer significant in the final model, and is therefore not included. This indicates that there is no credible evidence that performance on the job is dependent on the time availability restrictions.

II.4.6.3. Matching Effects

Despite gathering considerable interest, there is very little research on the influence of matching on response propensity in the literature, mainly due to data restrictions. The strongest evidence is available in the study by Moorman et al. (1999), but is restricted to descriptive statistics. This study compares the cooperation rates for a case-control study of breast cancer for sample members that were allocated an interviewer of the same ethnicity with those sample members discordant with their interviewer on ethnicity. For this very particular research topic matching seems to positively influence cooperation rates. While Durrant et al. (2010) report some evidence of sex and educational

matching effects, the overall effect of a variable is not considered, and low significance levels (10% significance level) are reported.

In this study the inclusion of various matching effects on the full model is attempted, but none of these are found to be significant. Some attempted matching effects follow the normal mathematical convention for interactions, while others are included through dummy variables indicating whether there is a match between the wave 8 interviewer and the respondent (as specified in wave 7) on specific criteria. These criteria include sex, age, and respondent education and interviewer grade/experience. While recognising that the wave 7 respondent and the wave 8 interviewer are time discordant, the decision to focus on wave 8 interviewer and the relative stability of the 'mother figure' respondent across waves justifies this choice. The non-significance of such terms indicates that there is no evidence of reduced refusals for cases where the interviewer is similar to the respondent on socio-demographic variables.

A different type of interaction effect is considered in the study by Lynn et al. (2013), which found some indication of variation in the effect of interviewer change by respondent age and interviewer age. Interaction effects for the interviewer change variable and respondent, household or interviewer characteristics are attempted to identify whether the effect of a change in interviewer across waves on nonresponse varies by respondent or interviewer characteristic. In this paper all variables included in the model have been considered for this interaction effect. The results do not indicate any differences in the effect of interviewer change between categories of respondent or interviewer variables. The results on interaction effects are interpreted within the context of the possibility that the actual variables which show a significant interaction with interviewer change may not have been considered, owing to the fact that their main effect is not significant in the model or to data restrictions.

II.5. Conclusions

This paper explores cross-classified and multiple membership multilevel model specifications to account for area and interviewer effects on wave nonresponse in longitudinal surveys. A cross-classified model is identified as the appropriate method for distinguishing between area and interviewer effects in the case of partial interpenetration, otherwise known as cross-classification, sometimes present in surveys. Cross-classified and multiple membership specifications are considered to account for the various interviewers allocated to a particular case across waves. The analysis of wave 8 nonresponse for the UK Family and Children Survey serves as an example to illustrate the methods proposed.

The main results from this application are as follows. The final random effect specification identified for this dataset is a two-level hierarchical model with a random effect for the current interviewer. Area effects are not significant after controlling for interviewer and household level effects in a cross-classified model, supporting findings by Campanelli and O'Muircheartaigh (1999) and Durrant et al. (2010). The non-significance of the cross-classified area effect in comparison to the significant area effect in a two-level hierarchical model either suggests that there is insufficient interpenetration to correctly disentangle the two random effects, or that area effects are simply aggregated interviewer effects. To the extent that interviewers work in restricted geographical areas apparent variations in response rates across areas may simply represent variations in response rates across interviewers, with area classifications acting as a rough proxy measure for interviewer classifications. One would expect that area effects on nonresponse, signifying variations across communities in privacy and safety concerns, as well as attitudes towards cooperation, apply mainly to the first interview request. Once one successful interview has been secured with a particular individual or household, such concerns would most likely not be of relevance anymore. This result has been confirmed in the study by Campanelli and O'Muircheartaigh (1999) which finds no significant area effect on nonresponse, refusal and non-contact at the second wave of the British Household Panel Study when using data with a quasi-randomised design.

Alternatively, the physical, social and cultural spatial divisions related with nonresponse patterns may not match the PSU classification. The possibility of a significant area effect for a different area classification cannot be completely ruled out, raising questions on the validity of the obtained results in the case of an omitted crossed-factor (Luo & Kwok, 2009). The unstable estimates obtained from the cross-classified model controlling for both interviewers at wave 7 and 8 but excluding area effects suggest that the assumption of independent interviewer effects is erroneous. Alternatively, the percentage of cases with a change of interviewer was insufficient.

The results for the multiple membership models with various weight specifications indicates that the best model fit pertains to the model allocating the highest weight to the current wave interviewer. These findings indicate that the current wave interviewer seems to have the greatest impact on current wave nonresponse for later waves of a longitudinal study. They are in contrast with earlier findings by Pickery and Loosveldt (2002) who report that the first interviewer has the greatest influence. They investigated, however, interviewer effects at the beginning of a longitudinal study, analysing wave 1 and 2 interviewers, and used a cross-classified multilevel model specification. For our example, the multiple membership model does not seem to provide an improvement on the simpler 2-level hierarchical model accounting only for the current wave interviewer random effect.

The results from the final model confirm previous findings on the positive relationship between wave participation and interviewer experience, grade and continuity variables, highlighting for example the importance of retaining experienced interviewers within the agency. The non-random nature of interviewer change in observational studies, however, hinders the interpretation of the effect of interviewer continuity on response (for an investigation of this effect using experimental data see Lynn et al., 2013). The current study also sheds light on the need for further data on the work history of interviewers, as results indicate that experience in other interviewing modes and survey areas may be detrimental in obtaining cooperation in face-to-face interviewing in social surveys.

This study suggests that for later waves of general household longitudinal surveys, which are not of a particularly sensitive nature, sampled members do not seem to base their decision to cooperate on the demographic characteristics of the interviewer, and neither are they affected by whether they are discordant with the interviewer on such characteristics. As information on the ethnicity of interviewers was not available, the benefit of matching interviewers and households on ethnicity has not been explored here. In high-crime neighbourhoods, where there may be known frictions between ethnicity groups, households being contacted by an interviewer of the same ethnicity may feel less uneasy about safety concerns, and as Moorman et al. (1999) point out, these may also perceive the survey to be more relevant to their race group.

The results do not provide much support for the hypothesis that interviewer personality traits are important predictors of wave nonresponse. Despite being categorised as skills items, some of these items seem to be representing very specific personality characteristics rather than learnable behaviours. Although some of these items are found to be significant, the overall picture of the personality profile and skills set of the most successful interviewer in terms of nonresponse is not clear or coherent. The non-significance of these variables may, however, simply reflect an inadequate construct of personality and skill, or possibly a conscious decision taken by some interviewers to answer the questions in a favourable way, leading to distorted personality and skills assessments. Even if there is a relationship between a household's propensity to respond and the interviewer's personality, it may be too weak or complex to identify, and may therefore be of limited use in guiding interviewer recruitment and training. Cross-level interactions for personality traits have not been explored due to a lack of respondent personality information, and therefore the possibility of higher response rates for sampled members approached by interviewers of a similar personality typology cannot be ruled out.

It is important to remember that this study focuses on nonresponse at a later wave in the life of a longitudinal study. Results from this study may not apply for earlier waves. One might expect marked differences in the influence

of interviewers and areas on nonresponse across various stages of a longitudinal study. For example, area and interviewers effects on nonresponse are expected to be greater at initial waves, and if the second wave response is analysed the relative influence of the first wave interviewer is expected to be greater than that of previous wave interviewers for subsequent waves.

It is important to consider how the cross-classification present in the FACS dataset came about. The survey administrators at NatCen explained that interviewers are allocated work within a geographic pool of a few primary sampling areas close to their area of residence. Interviewer case allocation is not random, but based on practical considerations. For example, if the previous interviewer leaves the agency an alternative interviewer is recruited to take over the caseload. Therefore, to the extent that interviewers are not matched to cases on the basis of their performance record or experience and if there is no substantial interaction effect between the interviewer area of residence and the area provenance of the cases allocated to a particular interviewer then minimal confounding in interviewer and area effects is expected. If these assumptions hold, then the cross-classification observed can be considered to provide partial interpenetration. However, if these assumptions do not hold then disentangling interviewer and area effects may be problematic.

III. The Effect of Sample Size and Level of Interpenetration on Inference from Cross-classified Multilevel Logistic Regression Models (Paper 2)

III.1. Introduction

In survey methodology, 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 practice in survey designs owing to restrictions in field administration capabilities and survey costs. 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 cases in one area may be designated to more than one interviewer. 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 in cases where there is partial interpenetration (Von Sanden, 2004). However, 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., 2010) 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 under difficult design conditions have not yet been explored.

III.2. Study Aims

This study examines the implications of various practical limitations in the assignment of cases from different areas to interviewers within various scenarios through a simulation study. The implications are assessed in terms of the following measures: the percentage relative bias and the standard error of the area and interviewer variance estimators, the asymptotic Wald 95% confidence interval coverage, the correlation of the two variance estimators and the power of significance tests. 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 starting values for the other factors represent realistic values, making the simulation results relevant to survey practice. The study will also examine 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 are specific, and restricted to survey design and the exploration of interviewer effects on nonresponse, the same problem of identifiability may arise in other settings. Other survey design applications may consider the variation in the response to specific binary questionnaire items attributable to interviewers, in an attempt to quantify any interviewer influence on responses. For some items, area differences may be well documented, such as for questionnaire items asking about engagement in anti-social behaviour or views on the acceptability of various social trends. Other applications with similar design issues can also be envisaged.

Here an example of a similar design in a different subject matter, incorporating higher-level classifications other than areas and interviewers, is considered briefly. 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. It is important to note that unless physiotherapists are recruited separately from the national health system for this study the classification structure of the data may lead to confounding. Real work allocations may reflect unmeasured population density or resources limitations, which in turn may be related to the health outcome of the patient

III.3. Background

III.3.1. Two-level Hierarchical Models

The impact of various factors on the quality of model estimates may be assessed through simulation studies. Various studies have considered the impact of a number of factors on both fixed and random parameter estimates in two-level models for continuous outcome variables. Paccagnella (2011) summarises these results as follows: for a fixed sample size, increasing the number of clusters – rather than the number of units per cluster – yields more accurate, that is, less biased, parameter estimates and standard errors.

Estimates of random parameters are more prone to show non-negligible bias for small sample sizes than fixed parameter estimates, and more so when the intra-class correlation coefficient is high. Underestimation of standard errors is more pronounced for random parameter estimates, though still noticeable for fixed parameter estimates, especially when the number of groups is small.

Very few studies have explored the properties of estimators for binary outcome variables. These mainly include the studies by Rodriguez and Goldman (1995), Paccagnella (2011), Moineddin et al. (2007), and Theall et al. (2010). Both Moineddin et al. (2007) and Paccagnella (2011) use the NLMIXED procedure in SAS software to estimate the models, which calculates a maximisation of an approximated likelihood, integrating over the random effects. Theall et al. (2010) fit the models using the PROC GLIMMIX procedure in SAS software with restricted maximum likelihood estimation. Below, the results from these studies, on the impact of various factors – including the prevalence of the outcome, sample sizes and intra-cluster correlation (ICC) – on the point estimates and their standard errors is reviewed. The focus in this literature review is on binary outcome models. Other studies, mentioned for comparison purposes, refer to continuous outcome models.

III.3.1.1. Effect of Low Prevalence Outcome

Moineddin et al. (2007) show that scenarios with a very low probability of a successful outcome, e.g. 0.1 overall probability, compared to scenarios with moderate probabilities, e.g. 0.34 and 0.45 overall probabilities, show significantly higher bias for both fixed and random parameter estimates, and lower rates of model convergence. While the overall effect – considering four fixed effects, one random intercept and one random slope – of outcome probability on the Wald 95% confidence interval coverage rate is not significant, some differences in these rates can be observed for the two random effects parameters, with the lowest coverage rates obtained for the smallest levels of outcome prevalence. Similarly, the rate of model convergence was lowest for an overall probability of 0.1, with practically no difference between the 0.34 and the 0.45 overall probabilities.

III.3.1.2. Effect of Sample Size

III.3.1.2.1. Effect of Sample Size on Point Estimates

Paccagnella (2011) finds fixed effect coefficients, with the exception of contextual variables, most especially dichotomous contextual variables, always to be unbiased even for small sample sizes, with the smallest sample size considered consisting of 650 cases (10 groups). However, only the random intercept estimate is consistently underestimated for all sample sizes, including the largest sample size of 22,750 (350 groups). Bias reductions can be noticed as the number of groups, and consequently sample size increases up to 70 groups (4550 cases). Thereafter, any sample increases do not translate into improvements in the estimates' accuracy. In contrast, Moineddin et al. (2007) show that with small group sizes and a small number of groups, both fixed and random parameter estimates are biased. However, in agreement with the previous study reviewed, random effects parameter estimates show the largest biases. For fixed effect parameters bias is reduced to 1% or less for data with 100 groups of size 30. In the case of random effects estimates, recorded biases never reach 1% or lower, even for large sample sizes of 100 groups of size 50, at which point they reach up to 4%. Larger biases are recorded for the random intercept rather than the random slope estimates. For group sizes of 30 or more, irrespective of the number of groups, the random intercept and random slope parameters are underestimated.

Maas and Hox (2005) find no significant difference in bias across either number of groups or group sizes for their main simulation study. In contrast, in their additional simulation study including data with only 10 groups of size 5, substantial positive bias reaching 25% is recorded for group-level variances. The authors run this additional simulation to test Snijders and Bosker's (1999) statement that ten groups are the minimum adequate number for use in multilevel models. Maas and Hox (2005) conclude that having such a small group sample size is insufficient. A possible criticism of this study is that a larger group size (yielding larger total sample sizes) could have been considered for a fixed group sample size of 10. These may have provided more insightful results.

Theall et al. (2010) explore the influence of small group sizes for varying numbers of groups on the parameter estimates and their standard errors for both a continuous and binary outcome. The authors consider twenty five factor combinations. Scenarios with 90%, 75%, 50%, 25% and 10% of areas including 1, 2, 3, 4 and 5 individuals are considered. When all 459 areas are sampled for models including no fixed effects and models including individual and contextual variables, there are very slight differences in both the fixed and random parameter point estimates across these scenarios. As the number of areas is decreased from the total number of areas – 459 areas – to the minimum number of areas sampled in this simulation – 30 areas, the point-estimate for the higher-level parameter increases. For the results presented in the paper, specifically the results for scenarios with 90% of areas having only 2 individuals sampled within them, the greatest increase in the area variance is observed between 50 groups and 30 groups. This inflation of the variance estimate for small sample sizes is similar to what is observed in Maas and Hox (2005) for their small-sample data. Rodriguez and Goldman (1995) also find some evidence of the influence of cluster size upon the accuracy of point estimates. The authors find much higher negative bias for the variance parameter estimate for the family-level (which has a very small cluster size), corresponding with level two, compared to the community-level, corresponding with level three.

III.3.1.2.2. Effect of Sample Size on Standard Error Estimates

Moineddin et al. (2007) find that a larger number of groups results in higher 95% confidence interval coverage for both the random intercept parameter and the random slope parameter, but has no effect for fixed effect parameters. A larger group size results in significantly lower coverage only for the random slope parameter. This lower coverage is noticeable for the group size of 30 or the group size of 50 in comparison to the group size of 5. Coverage is close to the expected 95% mark for fixed effects parameters, while being lower than this nominal value for random effects parameters for all simulation conditions. This result indicates that standard error estimates for the random effects variance is underestimated. Paccagnella (2011) finds that the coverage rate is

just below nominal level for the overall intercept when only 30 groups, and sometimes 50 groups, are included, and at nominal level for all number of groups (10, 30, 50, 70, 100, 150 and 350 groups) for all level-1 fixed effects parameters, and for level-2 fixed effects once at least 50 groups are included. For random effects parameters, data with 70 groups (4550 cases) still shows coverage rates lower than 92%. While higher number of groups, and consequently sample sizes, do increase the coverage rate to some extent, the rate of increase is not sufficient, such that for 350 groups the average coverage rate across different conditions is around 93%.

Maas and Hox (2005) find similar results when considering a continuous outcome. Variance parameters always show lower coverage rates – implying underestimation of the length of the confidence interval – than the fixed effects parameters, which are generally close to the 95% nominal rate. A highly significant and substantial increase in the coverage rate across number of groups is observed for the random effects parameters. The coverage rates for the random intercept and random slope parameters for 30 groups are 91.1% and 91.2%, and these are increased to 94% and 94.3% when including 100 groups. For all fixed effects parameters, the coverage rate does not fall below 93.6% for the smallest group size, which is not too far off the nominal 95% rate. An increase from 5 to 30 in the group size generally seems to increase coverage, but an increase to 50 does not seem to benefit the coverage properties of the parameter's confidence intervals. Also, group size seems to have a smaller impact on coverage than the number of groups. In their additional simulation study, Maas and Hox (2005) find that for small sample sizes the standard error are underestimated for both fixed and random effects parameters, with coverage rates reaching 90.3% and 69.6% for fixed and random effect parameters respectively.

Theall et al. (2010) find that standard errors are inflated and confidence intervals become less precise for samples with a higher percentage of areas of small group size ($n \leq 5$). The authors however argue that this trend may simply reflect the smaller sample sizes for scenarios with a higher number of areas within which a small group size was sampled. The increase in the standard errors of both fixed and random parameter estimates with a smaller number of

groups were more substantial. The authors argue that they cannot categorically explain this pattern in terms of the smaller sample sizes due to the lack of a standard error formula for restricted maximum likelihood estimators. What is definitively concluded is that even with a group number which is usually considered sufficient in the literature – for example, 100 groups – if the majority of groups have very small group sizes – for example just 1 or 2 individuals per group – then the higher-level random effects and any contextual fixed effects may be found insignificant simply because of type II errors. This applies even in scenarios with a relatively high ICC. Similarly, Rodriguez and Goldman (1995) find inflated standard errors for the family-level (level two) variance estimate. They explain it in terms of the small number of units within each cluster at this level.

III.3.1.2.3. Effect of Sample Size on Model Convergence

The convergence rate in Moineddin et al. (2007) and Paccagnella (2011) indicates the percentage of times parameter convergence for a simulated dataset is not obtained by 1000 and 200 iterations respectively. Moineddin et al. (2007) find that the rate of model convergence increased for both increases in group sizes and in the number of groups. Similarly, Paccagnella (2011) shows that it is only data with only 10 groups that yields serious convergence problems. For data with 30 groups, non-convergence is only a problem with the smallest ICC. For all simulation conditions non-convergence is reduced to zero for data with at least 50 groups.

III.3.1.3. Effect of ICC

III.3.1.3.1. Effect of ICC on Point Estimates

Goldstein (2011) explains that the value of the ICC may also influence the estimates' accuracy. In the simulation study by Moineddin et al. (2007), the overall relative bias differs significantly by ICC values – set as 0.04, 0.17 and 0.38 – only for the random intercept, showing higher bias for lower ICC values. On the other hand, the random slope estimates and fixed effects estimates for

data of different ICC values do not show statistically significant bias differences. Maas and Hox (2005) find that in spite of the fact that the largest bias corresponds to the scenario showing the smallest sample sizes and the highest ICC, there are no significant differences in bias across ICC values. Paccagnella (2011) finds no effect of the ICC value – specified as 0.071, 0.304 and 0.655 – on the relative bias of all estimates.

III.3.1.3.2. Effect of ICC on Standard Error Estimates

In Moineddin et al. (2007), while varying ICC values show no effect on the Wald 95% confidence interval coverage rate for all fixed effect parameters and the random slope parameter, the random intercept parameter shows a trend of more accurate coverage rates for higher ICC values. Maas and Hox (2005) find no significant difference in the coverage rates across the different intra-class correlations specified: 0.1, 0.2 and 0.3. Similarly, Paccagnella (2011) finds no consistent effect of the ICC value on the coverage rates. Maas and Hox (2005) find similar results for the continuous case; coverage rates for the various parameters remain stable for the different ICC values considered: 0.1, 0.2 and 0.3.

III.3.1.3.3. Effect of ICC on Model Convergence

In Moineddin et al. (2007), convergence problems are most pronounced for data with group size 5, and this applies for all three simulations conditions having this group size but varying in the number of groups: 30, 50 and 100. For simulation conditions specified above the improvement in the rate of convergence for increasing ICC values is pronounced, more so for data with the smallest number of groups (30 groups). For the sample including 30 groups of size 5, the convergence rate increases from 56% for an ICC of 0.04 to 68% for an ICC of 0.17, to 75% for an ICC of 0.38. Similarly, the results in Paccagnella (2011) show non-convergence problems for the simulation conditions including only 10 groups. This is only substantial for the lowest ICC of 0.071, showing an 86% convergence rate for a sample size of 650 belonging

to 10 groups, whereas higher ICC data show non-convergence rates of less than 1%. In Mass and Hox (2005), models for all data scenarios converged.

III.3.2. 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 been considered yet. 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, as the authors consider Markov Chain Monte Carlo (MCMC) estimation in MLwiN too inefficient in terms of computational time. The most flexible template of 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.

In the MLPowSim manual the example considered in the cross-classified data section is an educational one, 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, supporting the earlier findings in for the two-level case, as reviewed in Scherbaum and Ferreter (2009). Also, adding further cases per higher-level grouping only benefits power calculations up to a threshold number of cases. Although this software offers a great template for sample size calculations for specific power requirements, no analysis of estimate accuracy is possible. To the extent that the percentage variation in nonresponse attributable to interviewers needs to be estimated accurately, this calculation is particularly relevant for this field of research and cannot be ignored.

III.4. Methodology

This section presents the details of the simulation design. The first section presents the cross-classified multilevel logistic regression model being fitted to the simulated data. In the next section, the process by which the data is generated is explained in detail. The various simulation scenarios and the design factor values considered are then specified. Next, the stored quantities from each fitted model are listed and the properties calculated from these stored quantities – including the rationale for considering each measure and the equations used for their calculation – are presented.

III.4.1. Simulation Model

The following model is used:

$$\text{logit}\left(p_{i(js)}\right) = \eta_{i(js)} = \beta_0 + u_j + v_s \qquad \text{(III. 1)}$$

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 analysis of the simulated datasets is carried out, that is, the models are fitted and parameters estimated, 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. Different burin-in lengths were attempted for different scenarios to identify the appropriate burn-in length to avoid undue influence from the starting values (Gelman et al., 2004). The Brooks-Draper and Raftery-Lewis diagnostics were checked for a selection of scenarios for different iteration lengths to determine the most time efficient length for accurate point estimates and 95% credible intervals. Initial values for parameters are obtained by making use of the second order penalised quasi-likelihood (PQL) estimation method. Due to the computational power and efficiency requirements of MCMC estimation, and the large number of models estimated, the IRIDIS High Performance Computing Facility, and the associated

support services at the University of Southampton have been required to complete the model estimation work.

III.4.2. Data Generating Procedure

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

$$\beta_0 = \log_e \frac{\pi}{1-\pi}. \qquad \text{(III. 2)}$$

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:

$$p_{i(js)} = \frac{\exp(\eta_{i(js)})}{1+\exp(\eta_{i(js)})}. \qquad \text{(III. 3)}$$

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 each scenario 1000 datasets are generated using R Version 2.11.1. Ritter et al. (1996) explain that the standard error of a mean is inversely proportional to the square root of the number of runs. The chosen number of datasets – 1000 runs – should be sufficient to produce stable estimates of the properties (presented in Section 4.5) while keeping running time and memory space requirements manageable.

For scenarios which vary only in the interviewer case allocation the same set of 1000 cluster-specific random effects is used. So, for a specific overall probability of response, overall sample size, number of areas and interviewers, and interviewer and area variances, the same interviewer and area residuals are

used across scenarios simply varying in terms of the interviewer case allocation scheme. 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. Therefore the 1000 simulations represent 1000 samples of a specific number of interviewers (medium scenario = 240 interviewers) sampled from an infinite population of interviewers. Then for each of the 1000 samples, all possible realistic scenarios of interviewer-area combinations are considered.

The procedure used for generating the data for one specific scenario is presented in Appendix B. The same programming code, with the appropriate changes to the factor values and allocation schemes, can be used for other scenarios.

III.4.3. Simulation Scenarios and Factors

To explore the properties of estimators, a simulation experiment is carried using a factorial design. The simulated scenarios vary in the following factors: the overall sample size, N, the number of interviewers and areas, N^I and N^A, and by consequence the number of cases per interviewer and per area, the level of cross-classification between interviewer and area allocations, the higher-level variance, and the overall probability of the outcome variable π. In this particular application relating to sampled cases allocated to an interviewer and residing within a particular primary sampling area, the higher-level variance σ_z^2 is divided in two parts – the interviewer-level variance σ_u^2 and the area-level variance σ_v^2. These variances will be altered one at a time, and also simultaneously, to explore changes in the estimators' properties arising from changes in ICC values.

The choice of the values for the various factors reflects realistic representations of general household survey scenarios. For some factors three different values are considered for each factor, representing low, medium and high scenario values. The medium scenario values are similar to values observed in available studies, which is used in this study as a realistic starting

point. In various studies, the same number of primary sampling units is maintained across waves. Consequently, 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 this 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 medium 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 and the 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 in the interviewers working in neighbouring areas – on the properties of the estimators and test statistic for the medium scenario factors is analysed. The number of areas each interviewer works in will be allowed to vary from 1 to 6. Imagining a situation where for a national survey an interviewer is asked to work in more than 6 primary sampling units is quite unrealistic, and therefore will not be considered here.

The diagrams below show the area-interviewer allocations for a few 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. Here, the simplest to the most geographically dispersed example considered are presented in this order. The simplest case – CASE 1 – is where two interviewers work in each area, with each interviewer working only in one area. 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. For the purpose of estimation, a cross-classified model can still be fitted to this data since MCMC estimation does not require the data structure to be perfectly identified, since unlike for the Iterative Generalised Least Squares (IGLS) which requires the global block diagonal matrix to be defined the MCMC method simply treats each unique classification structure as a random additive term (Browne et al., 2001). To check this equivalence in the

specification a few of the models for CASE 1 allocation schemes were also run using purely hierarchical 3-level models. The results obtained were equivalent to those obtained using the cross-classified model.

CASE 1

Area	Interviewers											
1	1	2										
2			3	4								
3					5	6						
4							7	8				
5									9	10		
6											11	12

Next, an interviewer can work in two areas, with four interviewers working in each area. Three possible scenarios may 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). Or finally, pairs of interviewers are always allocated together, with each particular pair never occurring twice with another pair (CASE 2C).

CASE 2A

Area	Interviewers											
1	1	2	3	4								
2	1	2	3	4								
3					5	6	7	8				
4					5	6	7	8				
5									9	10	11	12
6									9	10	11	12

CASE 2B

Area	Interviewers											
1	240	1	2	3								
2		1	2	3	4							
3					4	5	6	7				
4						5	6	7	8			
5									8	9	10	11
6										9	10	11	12

CASE 2C												
Area	Interviewers											
1	239	240	1	2								
2			1	2	3	4						
3					3	4	5	6				
4							5	6	7	8		
5									7	8	9	10
6											9	10	11	12

For cases where interviewers work in three areas, each area includes six different interviewers. The different allocation possibilities for this specification are depicted below. In the first case there is a group of six interviewers who always work together, and who do so in three different areas (CASE 3A). In the next case a group of five interviewers always work together, and for two instances out of three a group of six interviewers are maintained across neighbourhoods (CASE 3B). Then, groups of five interviewers are maintained in two instances out of three (CASE 3C). There are some other overlaps, but with small groups of interviewers compared to the previous case. CASE 3D and 3E provide two different possibilities for maximum overlaps of four interviewers. In CASE 3E there is an overlap of four interviewers in two out of three areas. In CASE 3F overlaps are restricted to three interviewers across all three areas each interviewer is working in. Finally, in CASE 3H overlaps are restricted to two interviewers across all three areas each interviewer is working in. Each pair of interviewers does not work with another pair more than once, and therefore there are no other overlaps.

CASE 3A												
Area	Interviewers											
1	1	2	3	4	5	6						
2	1	2	3	4	5	6						
3	1	2	3	4	5	6						
4							7	8	9	10	11	12
5							7	8	9	10	11	12
6							7	8	9	10	11	12

CASE 3B

Area	Interviewers
1	240 1 2 3 4 5
2	1 2 3 4 5 6
3	1 2 3 4 5 6
4	6 7 8 9 10 11
5	7 8 9 10 11 12
6	7 8 9 10 11 12

CASE 3C

Area	Interviewers
1	236 237 238 239 240 1
2	237 238 239 240 1 2
3	1 2 3 4 5 6
4	2 3 4 5 6 7
5	3 4 5 6 7 8
6	7 8 9 10 11 12

CASE 3D

Area	Interviewers
1	237 238 239 240 1 2
2	237 238 239 240 3 4
3	1 2 3 4 5 6
4	1 2 3 4 7 8
5	5 6 7 8 9 10
6	5 6 7 8 11 12

CASE 3E

Area	Interviewers
1	237 238 239 240 1 2
2	239 240 1 2 3 4
3	1 2 3 4 5 6
4	3 4 5 6 7 8
5	5 6 7 8 9 10
6	7 8 9 10 11 12

CASE 3F

Area	Interviewers
1	238 239 240 1 2 3
2	1 2 3 4 5 6
3	1 2 3 7 8 9
4	4 5 6 7 8 9
5	4 5 6 10 11 12
6	7 8 9 10 11 12

CASE 3H

Area	Interviewers
1	1 2 3 4 5 6
2	1 2 7 8 9 10
3	1 2 11 12 13 14
4	3 4 7 8 11 12
5	3 4 9 10 13 14
6	5 6 7 8 13 14

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. Below are the cases considered when each interviewer works in four areas, and cases within each area are allocated to eight different interviewers (CASE 4A, 4B & 4C).

CASE 4A

Area	Interviewers
1	1 2 3 4 5 6 7 8
2	1 2 3 4 5 6 7 8
3	1 2 3 4 5 6 7 8
4	1 2 3 4 5 6 7 8
5	9 10 11 12 13 14 15 16
6	9 10 11 12 13 14 15 16

CASE 4B

Area	Interviewers
1	1 2 3 4 5 6 7 8
2	2 3 4 5 6 7 8 9
3	3 4 5 6 7 8 9 10
4	4 5 6 7 8 9 10 11
5	9 10 11 12 13 14 15 16
6	10 11 12 13 14 15 16 17

CASE 4C

Area	Interviewers
1	239 240 1 2 3 4 5 6
2	1 2 3 4 5 6 7 8
3	3 4 5 6 7 8 9 10
4	5 6 7 8 9 10 11 12
5	7 8 9 10 11 12 13 14
6	9 10 11 12 13 14 15 16

Below the allocation schemes where each interviewer works in five areas, with each area including ten interviewers, are presented (CASE 5A, 5B & 5C).

CASE 5A

Area	Interviewers
1	1 2 3 4 5 6 7 8 9 10
2	1 2 3 4 5 6 7 8 9 10
3	1 2 3 4 5 6 7 8 9 10
4	1 2 3 4 5 6 7 8 9 10
5	1 2 3 4 5 6 7 8 9 10
6	11 12 13 14 15 16 17 18 19 20

CASE 5B

Area	Interviewers
1	1 2 3 4 5 6 7 8 9 10
2	2 3 4 5 6 7 8 9 10 11
3	3 4 5 6 7 8 9 10 11 12
4	4 5 6 7 8 9 10 11 12 13
5	5 6 7 8 9 10 11 12 13 14
6	11 12 13 14 15 16 17 18 19 20

CASE 5C

Area	Interviewers
1	1 2 3 4 5 6 7 8 9 10
2	3 4 5 6 7 8 9 10 11 12
3	5 6 7 8 9 10 11 12 13 14
4	7 8 9 10 11 12 13 14 15 16
5	9 10 11 12 13 14 15 16 17 18
6	11 12 13 14 15 16 17 18 19 20

Finally, the allocation schemes below represent scenarios where each interviewer works in six areas, with each area having twelve interviewers working within it (CASE 6A, 6B & 6C).

CASE 6A

Area	Interviewers
1	1 2 3 4 5 6 7 8 9 10 11 12
2	1 2 3 4 5 6 7 8 9 10 11 12
3	1 2 3 4 5 6 7 8 9 10 11 12
4	1 2 3 4 5 6 7 8 9 10 11 12
5	1 2 3 4 5 6 7 8 9 10 11 12
6	1 2 3 4 5 6 7 8 9 10 11 12

CASE 6B

Area	Interviewers																	
1	1	2	3	4	5	6	7	8	9	10	11	12						
2		2	3	4	5	6	7	8	9	10	11	12	13					
3			3	4	5	6	7	8	9	10	11	12	13	14				
4				4	5	6	7	8	9	10	11	12	13	14	15			
5					5	6	7	8	9	10	11	12	13	14	15	16		
6						6	7	8	9	10	11	12	13	14	15	16	17	

CASE 6C

A	Interviewers																					
1	1	2	3	4	5	6	7	8	9	10	11	12										
2			3	4	5	6	7	8	9	10	11	12	13	14								
3					5	6	7	8	9	10	11	12	13	14	15	16						
4							7	8	9	10	11	12	13	14	15	16	17	18				
5									9	10	11	12	13	14	15	16	17	18	19	20		
6											11	12	13	14	15	16	17	18	19	20	21	22

Due to computer power limitations and dependencies between factors – such that, for example, 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. Any changes are implemented for a select number of interviewer work allocation schemes (rather than attempting all schemes for every single factor change), for efficiency reasons. Table 1 outlines the medium values as well as the other values considered for each factor in the simulation study.

Table III.1: Factor Values for Medium and Other Scenarios

Factor	Medium	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

The analysis for the initial medium 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 medium 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 medium 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 below, denoted as CASEi, where i represents the number of areas each interviewer works in, are therefore comparable to the allocation schemes CASEiA outlined above.

CASE 1

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

CASE 2

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

CASE 3

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

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

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

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					

This paper will not consider the impact of different estimation methods on the properties of the estimated parameters. Neither is the impact of different distributional assumptions for the higher-level variances considered. The focus on the random effects parameters is justified by the fact that consistently across various studies, looking at 2-level models with either continuous or binary outcomes, random effects estimates and their respective standard errors were more inaccurate than fixed effects estimates. Moreover, the primary aim of this paper is to identify how well the variances of the two higher-level classifications are estimated for realistic allocation of cases to interviewers.

III.4.4. Stored Quantities for each Model

For each simulation, the parameter estimates, the standard errors and the 95% confidence intervals are obtained. Two confidence intervals are obtained: one 95% confidence interval is obtained using the asymptotic normal distribution; the other reflects the credible 95% confidence interval obtained from the MCMC quantiles. The Bayesian interval provides an alternative to the maximum likelihood confidence interval, which is not reliant on the assumption of a normal sampling distribution. The variance-covariance matrix of the parameter estimates is also obtained. For each parameter of each model the effective sample size, the mean Monte Carlo standard error, the Brooks-Draper diagnostic, and the lower and upper bound of the Raftery-Lewis diagnostic are also obtained. For each model the Deviance Information Criterion and the time taken for the model to be estimated are recorded.

III.4.5. Properties of the Estimators and Test Statistic

The models are assessed in terms of various properties: the correlation of the two variance estimators, 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 for each dataset is calculated using the formula

$$\frac{1}{1000}\sum_{i=1}^{1000} corr_i\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right) = \frac{1}{1000}\sum_{i=1}^{1000}\frac{cov_i\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)}{\sqrt{var_i\left(\widehat{\sigma_u^2}\right)var_i\left(\widehat{\sigma_v^2}\right)}}. \qquad \text{(III. 4)}$$

'Good' estimators are expected to show no substantial correlation. High negative correlation values will 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 interviewer and area

variance 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 co-linearity 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 a parameter is calculated to determine the accuracy of a parameter estimator using the formula

$$\frac{1}{1000}\sum_{i=1}^{1000}\frac{\hat{\theta}_i - \theta}{\theta} * 100 \qquad \text{(III. 5)}$$

where $\hat{\theta}_i$ is the parameter estimate, θ is the true parameter value and i is the simulation number. 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 as substantial. This measure can also be calculated for the median MCMC credible interval value.

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 and the 95% credible confidence interval from the MCMC chains of the parameter estimate for each simulated dataset is recorded separately. Various authors, including Maas and Hox (2005) and Paccagnella (2011), explain that there are problems in using the standard approach to constructing the confidence intervals for random effects components. The standard approach assumes normality, while variances can only be positive. No study could be identified which uses the MCMC quantiles as an alternative for confidence intervals, which is advocated in this paper as a better measure. If the lower and upper bound of the Wald confidence interval (or the relevant quantiles of the MCMC chain) are either both less than or both greater than the true parameter value then the confidence interval does not cover the true parameter value. For those samples which include the true parameter a value of 1 on the coverage indicator is assigned, while the other samples are given a value of 0. To determine the coverage of the 95% confidence intervals the percentage of the (1000) datasets for which the confidence interval included

the true parameter value is calculated. The coverage rate is recorded for all simulation scenarios and compared with the expected 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. It is calculated using the following equation:

$$\frac{1}{1000}\sum_{i=1}^{1000} S.E._i\,(\hat{\theta}) \qquad (6).$$

While the accuracy of estimators, estimated by the percentage relative bias measure, may not vary much across the various scenarios, such that the mean estimate is not substantially different, there may still be substantial differences in the standard errors, and therefore in the precision of estimators.

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. This consists of dividing the coefficient estimate by its standard error and squaring that value. The corresponding p-value for this value is obtained from a chi-squared distribution. If the p-value is greater than 0.05 then the null hypothesis is not rejected. The proportion of datasets for which the null hypothesis is not rejected is subtracted from 1 to obtain the power of the 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 validity of 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).

The procedure for running the models, storing the output quantities and calculating the properties is specified in Appendix C. The same programming

code, with the appropriate changes to the factor values and file names, can be used for other scenarios.

III.5. Results

To reiterate what has been outlined in the methodology section, the medium scenario design has the following properties: 120 areas (48 cases/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. For each property, both the factors which seem to have no effect and those factors which have an impact on the quality 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 optimal results across different factor values. Under all simulation scenarios the test for β_0 obtains a power of 1. Accurate intercept estimates $\widehat{\beta_0}$ 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 will be discussed further in the relevant results section. The main points are summarised in bullet points at the end of each section.

III.5.1. Power of Test

The power of the Wald test to detect a significant effect can reach extremely low levels when interviewers are allocated work in only one area. In some scenarios the power is equal to 0 for CASE 1 allocation schemes. For CASE 2 allocation schemes the lowest power obtained for the scenarios considered in this study is 0.67, a major improvement over the power results for CASE 1. 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. For the medium scenario design 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 (Table 2, Columns 1 & 2). In fact it is only the test for the area random parameter σ_v^2 for the least sparse interviewer allocation (CASE 1) that yields a power not equal to 1, being 0.91.

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). For example, for σ_v^2 the power is equal to 0.91 for N=5670, 0.63 for N=2880 and 0.30 for N=1440 for CASE 1. Therefore, sparser interviewer allocation schemes are required to obtain similar high levels of power (Table 2, Columns 3-6). For all three sample sizes the greatest improvement in the power comes from increasing the number of areas per interviewer from one to two. For the 1440 sample size scenario a substantial increase in power can also be observed when increasing the number of areas per interviewers from two to three. However, further dispersion only yields very small gains, and the sparsest and least overlap interviewer allocation (CASE 6C) only obtains a power of 0.91 for σ_v^2 and 0.89 for σ_u^2.

Table III.2: Power of Wald Test at the 5% Significance Level by Sample Size and Interviewer Allocation

	Sample Size					
	5760		2880		1440	
IA	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
2[a]	1.00	1.00	0.96	0.98	0.77	0.81
2B	1.00	1.00	0.99	0.99	0.78	0.83
2C	1.00	1.00	0.99	1.00	0.79	0.84
3A	1.00	1.00	1.00	1.00	0.91	0.89
3B	1.00	1.00	1.00	1.00	0.85	0.86
3C	1.00	1.00	0.99	0.99	0.85	0.84
3D	1.00	1.00	1.00	0.99	0.86	0.86
3E	1.00	1.00	1.00	0.99	0.85	0.84
3F	1.00	1.00	1.00	1.00	0.87	0.86
3H	1.00	1.00	1.00	1.00	0.87	0.85
4A	1.00	1.00	1.00	1.00	0.88	0.86
4B	1.00	1.00	1.00	1.00	0.88	0.86
4C	1.00	1.00	1.00	1.00	0.89	0.88
5A	1.00	1.00	1.00	1.00	0.91	0.89
5B	1.00	1.00	1.00	1.00	0.89	0.90
5C	1.00	1.00	1.00	1.00	0.91	0.87
6A	1.00	1.00	1.00	1.00	0.92	0.88
6B	1.00	1.00	1.00	1.00	0.91	0.90
6C	1.00	1.00	1.00	1.00	0.91	0.89

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

For $N^I = N^A$ scenarios a similar pattern for the change in power with decreasing N is observed. In contrast to $N^I = 2\,N^A$ scenarios there is a substantial improvement in the power values when moving from CASE 3 to CASE 4 for the N=1440 cases of the $N^I = N^A$ scenarios. For the $N^I = 2\,N^A$ scenarios this further dispersion did not yield a gain in power. Therefore, the negative influence of small N on the power of the Wald test is greater for scenarios with an equal number of higher-level units (N^I=N^A).

Table III.3: Power of Wald Test at the 95% Confidence Level by Sample Size and Interviewer Allocation

	Overall Probability					
	0.7		0.8		0.9	
IA	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2
1	0.96	1.00	0.91	1.00	0.80	0.95
2A	1.00	1.00	1.00	1.00	1.00	0.99
2C	1.00	1.00	1.00	1.00	1.00	0.99
3A-6C	1.00	1.00	1.00	1.00	1.00	1.00

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 3 presents scenarios that vary only by the overall probability (π=0.7, 0.8, 0.9) with other factors kept constant at the medium scenario values. These results show that for CASE 1 higher overall probabilities result in lower power for the random parameters σ_v^2 and σ_u^2. The power for the interviewer random parameter σ_u^2 only decreases for the highest overall probability 0.9, whilst the power for the test for the area random parameter σ_v^2 decreases more rapidly for both overall probabilities 0.8 and 0.9 compared to the lowest overall probability 0.7. This difference in trend for the power of the Wald test for the two random parameters may be explained in terms of the higher N^I (240 interviewers) included in the sample compared to the N^A (120 areas). 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. For sparser interviewer allocation schemes there is no substantial difference in the power of the tests by overall probabilities.

The only difference in power across different values of σ_v^2 and σ_u^2 is observed for the CASE 1 allocation scheme. For the scenario with smaller area variance (σ_v^2=0.2) and medium scenario values for the other factors 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 medium 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. Again, this difference in the effect of the variance on the power of the test can be explained by the fact that $N^I=2N^A$ in these scenarios. The number of higher-level units mediates the effect of a lower ICC on the power of the tests for the random parameters.

Interestingly, for a specific area variance value higher power for the test of the area parameter is obtained when the interviewer variance is of a smaller effect size for CASE 1. Therefore, the power for σ_v^2 when σ_v^2=0.2 is 0.82 when σ_u^2=0.2 and 0.68 when σ_u^2=0.3. The power for σ_v^2 when σ_v^2=0.4 is 0.99 when σ_u^2=0.3 and 0.85 when σ_u^2=0.4. Similarly, the power for σ_v^2 when σ_v^2=0.3 is 0.99 when σ_u^2=0.2, 0.91 when σ_u^2=0.3 and 0.84 when σ_u^2=0.4.

Table III.4: Power of Wald Test at the 95% Confidence Level by Sample Size, Ratio of Interviewers to Areas and Interviewer Allocation

	$N^I=2N^A$						$N^I=N^A$					
	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

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 4). This difference is substantial for the CASE 1 allocation scheme for the 5760 and 2880 sample size scenarios and CASE 1-CASE 2 for the 1440 sample size scenario (Columns 1-6). On the other hand, $N^I=N^A$ scenarios do

not show this pattern in the power differences for the two random parameters (Columns 7–12). This result reflects the importance of having a large number of higher-level units to obtain good power.

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. Comparing the scenarios with different interviewers to area ratios for CASE 2–CASE 6 (interviewer case allocation schemes with at least two areas per interviewers) for the 5760 sample size scenarios, power is observed to be constant – with a value of 1 – for all allocation schemes (Table 4, Columns 1–2, 7–8). For the 2880 sample size scenarios any deviations are small, with a magnitude of 0.04 or lower (Table 4, Columns 3–4, 9–10). On the other hand, for the 1440 sample size scenarios power deviations smaller than 0.03 are obtained for allocation schemes CASE 4–CASE 6 (Table 4, Columns 5–6, 11–12). Four areas per interviewer are required for the scenario including 1440 cases with $N^I=N^A$ (30 areas allocated to 30 interviewers, σ_v^2=0.3, σ_u^2=0.3, π=0.8), compared to three areas per interviewer for the scenario including 1440 cases with $N^{ints}=2N^{areas}$ (30 areas allocated to 60 interviewers, σ_v^2=0.3, σ_u^2=0.3, π=0.8), for power to be greater than or equal to 0.85. The greatest decrease in power for $N^I=N^A$ in comparison to $N^I=2N^A$ scenarios arises for the CASE 1 allocation scheme. The one area per interviewer allocation scheme with $N^I=N^A$ scenarios for all three values of N considered (5760, 2880, 1440 cases) yield unacceptable power, with the highest power obtained for the largest N of 5760 cases being 0.07 and 0.08 for σ_v^2 and σ_u^2 respectively.

To summarise, the main points on the power of the Wald test are the following:

- For the medium scenario design the power is higher than 0.9 for both random parameters for all case allocation schemes.
- 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 go down to 0 for certain scenarios, whereas the lowest power value 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.

- Reduced interviewer overlap for a constant number of areas per interviewer does not improve the power.
- A decrease in the sample size results in lower power, especially for the allocation schemes with the least interviewer dispersion. The effect of sample size reduction on power is greater for scenarios with an equal number of higher-level units.
- There is some evidence that very high overall probabilities result in lower power for the least sparse interviewer allocation scheme. 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.
- The power for the area parameter is lower than that for the interviewer parameter when there is twice the number of interviewers as there are areas. No difference is observed for scenarios with equal numbers of interviewers and areas.
- Scenarios with equal numbers of interviewers and areas require more interviewer dispersion than the scenarios with twice the number of interviewers to areas to obtain the same level of power.

III.5.2. Correlation between Random Parameter Estimators

High negative correlations up to a value of -0.91 between the two random parameter estimators are obtained for all scenarios when interviewers are working solely in one area (Table 5, Row 1). For the CASE 2 allocation scheme the largest negative $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ obtained is -0.19. This value is observed for the scenario with the following factor specifications: 5760 cases, 120 interviewers, 120 areas, σ_v^2=0.3, σ_u^2=0.3, π=0.8. For the medium scenario design $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ is less than -0.1 (the threshold value being considered as

problematic) for all interviewer case allocation possibilities, except for the most restrictive interviewer allocation (CASE 1) which is -0.45 (Table 5, Column 1).

For scenarios with smaller N, but keeping constant all other factors, no substantial increases in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ can be observed (Table 5, Columns 1-3). Only a slight decrease in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ for the one area per interviewer allocation scheme from the 5760 and 2880 sample size scenarios to the 1440 sample size scenario is observed. For the $N^I=N^A$ scenarios with varying N but keeping other factors constant (σ_v^2=0.3, σ_u^2=0.3, π=0.8), $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ decreases from 0.91 to -0.83 to -0.69 for the 5760 cases, 2880 cases and 1440 cases for CASE 1 (Table 5, Columns 4-6, Row 1). A very small decrease in the negative correlation is also present for the next interviewer case allocation scheme CASE 2. Thereafter, for more sparse allocation schemes no substantial differences can be observed for different sample size scenarios in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$. Therefore, the effect of N on $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ 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.

Table 5 shows that $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ are higher for $N^I=N^A$ scenarios than for $N^I=2N^A$ scenarios for interviewer case allocation schemes CASE 1-CASE 4. To obtain a correlation of -0.1 or less for a sample size of 5760 cases, the scenario $N^I=2\,N^A$ requires at least two areas per interviewers, whereas the scenario $N^I=N^A$ requires at least four areas per interviewers. This indicates that a higher number of clusters yields better estimates than a larger cluster size for a constant N. An unequal number of clusters for the two classifications being considered may also help reduce identifiability problems in apportioning the higher-level variance.

Table III.5: $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ 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

The increase in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ with increasing overall probabilities (π=0.7, 0.8, 0.9) can be observed up to allocation CASE 3A (Table 6). The increase in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ 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 $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ is monotonic but not linear. With sparser interviewer allocation schemes the correlation values are minimal, and mostly positive, and vary only minimally across scenarios with different overall probabilities π. These results indicate that the effect of the overall probability on $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ is almost cancelled out once each interviewer is working in three areas.

Table III.6: $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ by Overall Probability and Interviewer Allocation

	Overall Probability		
IA	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 7 shows the $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ values for different scenarios varying by interviewer allocation and the variances σ_v^2 and σ_u^2. The analysis of the pattern of change in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ for different values of the variances σ_v^2 and σ_u^2 will be limited to the negative correlations, that is, CASE 1–CASE 3A interviewer allocation schemes. The results show that a consistent pattern of change in $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ emerges only when the area variance σ_v^2 is varied. This is true for interviewer allocation schemes CASE 1–CASE 3A. For higher values of σ_v^2 , and therefore higher ICC, the negative $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ decrease in size. No clear trend can be identified for varying the interviewer variance σ_u^2. Here the discrepancy in group sizes, with $N^I=2N^A$, suggests that the impact of the ICC (dependent on the variance) on $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ is mediated by the number of groups pertaining to the higher-level classification.

Table III.7: $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ by Area and Interviewer Variance and Interviewer Allocation

IA	$corr(\widehat{\sigma_u^2},\widehat{\sigma_v^2})$						
	σ_v^2=0.3 σ_u^2=0.3	σ_v^2=0.2 σ_u^2=0.2	σ_v^2=0.4 σ_u^2=0.4	σ_v^2=0.3 σ_u^2=0.4	σ_v^2=0.4 σ_u^2=0.3	σ_v^2=0.3 σ_u^2=0.2	σ_v^2=0.2 σ_u^2=0.3
1	-0.45	-0.51	-0.42	-0.48	-0.37	-0.27	-0.53
2A	-0.09	-0.12	-0.07	-0.09	-0.07	-0.09	-0.11
2C	-0.05	-0.10	-0.03	-0.05	-0.03	-0.06	-0.08
3A	-0.03	-0.04	-0.01	-0.02	-0.01	-0.02	-0.04

The first $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ column (highlighted in orange) represents the medium scenario design. The cells highlighted in red show increases in negative correlations, while cells highlighted in yellow show decreases in negative correlations, compared with the medium scenario design.
Constant factor values: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.8, $N^I=2N^A$

Lower negative $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ can be noticed for interviewer allocation schemes which have less overlap for the 2 areas per interviewer allocation schemes (Table 8). No clear pattern of varying $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ for different levels of interviewer overlap in interviewer allocation schemes can be noticed for sparser schemes of three or more areas per interviewer (CASE 3 – CASE 6). 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.

Table III.8: $corr\left(\widehat{\sigma_u^2}, \widehat{\sigma_v^2}\right)$ by Different Scenarios

	Scenarios										
IA	A	B	C	D	E	F	G	H	I	J	K
2A	−0.09	−0.11	−0.09	−0.12	−0.08	−0.12	−0.07	−0.09	−0.07	−0.09	−0.11
2B	−0.06	−0.07	−0.07								
2C	−0.05	−0.06	−0.06	−0.10	−0.04	−0.10	−0.03	−0.05	−0.03	−0.06	−0.08

A–K represent different scenarios as specified below:
B: N=2880, N^I=120, N^A=60, σ_v^2=0.3, σ_u^2=0.3, π=0.8; C: N=1440, N^I=60, N^A=30, σ_v^2=0.3, σ_u^2=0.3, π=0.8;
D: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.9; E: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.7;
F: N=5760, N^I=240, N^A=120, σ_v^2=0.2, σ_u^2=0.2, π=0.8; G: N=5760, N^I=240, N^A=120, σ_v^2=0.4, σ_u^2=0.4, π=0.8;
H: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.4, π=0.8; I: N=5760, N^I=240, N^A=120, σ_v^2=0.4, σ_u^2=0.3, π=0.8;
J: N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.2, π=0.8; K: N=5760, N^I=240, N^A=120, σ_v^2=0.2, σ_u^2=0.3, π=0.8

To summarise, the main points on the correlation between random parameter estimators are the following:

- Interviewer dispersion highly influences the correlation 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.
- No effect of sample size on correlation is observed for allocation schemes which allocate interviewers to at least two areas.
- Scenarios with equal numbers of areas and interviewers obtain higher negative correlations than scenarios with twice the number of interviewers to areas. 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. For allocation schemes with at least three areas per interviewer this effect 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 the correlation.
- 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.

III.5.3. Percentage Relative Bias of Variance Estimators

In most scenarios with N=5760, the relative percentage bias for the variance parameters estimators is around 1-3% once interviewers are allocated work in at least two areas (Table 9, Column 1). The relative percentage bias is much higher for interviewer allocation schemes which restrict the interviewer to working in one area (CASE 1). The bias for CASE2-6 fluctuate around within the range specified above, failing to show any systematic trend in bias reduction with further dispersion and less interviewer overlap. For $N^I=2N^A$ scenarios, for the least geographically sparse interviewer allocation (CASE 1) the area effect is always under-estimated (negative bias) (Table 9, Columns 1-3, Row 1), whilst the interviewer effect is over-estimated (positive bias) (Table 9, Columns 4-6, Row 1). 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 9, Columns 1-6, Rows 2-6). This again confirms the importance of group size for the accuracy of parameter estimators. A counterintuitive result is the larger biases for interviewer random parameter σ_u^2 compared to the area random parameter σ_v^2 obtained for CASE 1 (Table 9, Columns 1-6, Row 1).

Table III.9: 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

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 9, Column 12). The estimators of the $N^I=N^A$ (Table 9, Columns 7-12) scenarios obtain higher biases than the $N^I=2N^A$ scenarios (Table 9, Columns 1-6) in most cases for interviewer allocation schemes CASE2-CASE6. 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, for the 1440 total sample size the $N^I=N^A$ scenario yields consistently higher biases than the $N^I=2\ N^A$ 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.

As shown in Tables 10 and 11 there is no clear pattern for the change in the percentage relative mean bias of the variance parameter estimators by overall probability π and variances σ_v^2 and σ_u^2.

Table III.10: Percentage Relative Bias Mean Estimate by Overall Probabilities

IA	$\widehat{\sigma_v^2}$ π=0.7	$\widehat{\sigma_v^2}$ π=0.8	$\widehat{\sigma_v^2}$ π=0.9	$\widehat{\sigma_u^2}$ π=0.7	$\widehat{\sigma_u^2}$ π=0.8	$\widehat{\sigma_u^2}$ π=0.9
1	-3.34	-3.24	-4.42	5.41	6.80	6.87
2A	2.11	2.02	1.71	1.01	1.32	-0.40
2C	1.33	2.42	0.37	0.86	-0.56	0.30
3A	1.38	2.35	1.56	0.72	0.13	-0.48
3E	1.95	1.78	0.87	0.14	0.12	-0.35
3H	1.92	1.86	0.56	1.00	-0.66	0.16
4A	1.46	1.73	1.34	0.94	0.72	-0.50
4C	3.00	1.29	1.03	0.65	0.57	-0.88
5A	2.05	1.73	1.84	1.48	0.96	0.68
5C	2.34	2.29	2.12	0.69	-0.21	0.40
6A	0.52	1.08	1.19	1.00	0.74	-0.47
6C	1.30	1.69	0.68	1.47	0.21	0.68

The columns highlighted in orange represent the medium scenario design (N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.8, N^I=2N^A). The cells highlighted in red show increases in the absolute bias, while cells highlighted in yellow show decreases in absolute bias, compared with the medium scenario design (orange). The other scenarios maintain the same factors as the medium scenario design except for the overall probability as specified above in the table header

Table III.11: Relative Percentage Bias by Scenarios Varying in the Area and Interviewer Variances

IA	A: σ_v^2=0.3, σ_u^2=0.3		F: σ_v^2=0.2, σ_u^2=0.2		G: σ_v^2=0.4, σ_u^2=0.4		H: σ_v^2=0.3, σ_u^2=0.4		I: σ_v^2=0.4, σ_u^2=0.3		J: σ_v^2=0.3, σ_u^2=0.2		K: σ_v^2=0.2, σ_u^2=0.3	
	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$
1	−3.24	6.80	−5.93	9.54	−2.00	5.57	−4.72	7.55	−0.38	2.74	0.02	2.66	−10.87	7.82
2A	2.02	1.32	1.82	0.58	3.27	0.37	2.21	1.24	2.73	0.94	2.30	−0.24	−0.84	1.41
2C	2.42	−0.56	2.32	−0.08	1.39	0.93	1.98	0.61	2.50	0.65	2.30	−1.29	0.09	0.94
3A	2.35	0.13	2.64	1.17	1.91	0.94	1.72	0.08	1.76	1.50	1.76	−1.75	0.26	0.92
3E	1.78	0.12	1.59	−1.44	2.66	0.75	1.98	0.87	2.79	−0.60	2.08	0.41	−0.22	1.05
3H	1.86	−0.66	2.44	1.20	2.09	0.61	1.15	0.67	2.07	−0.22	3.04	−1.84	−0.54	1.13
4A	1.73	0.72	1.63	0.26	2.17	2.01	3.04	0.65	2.30	0.46	1.54	−0.25	0.64	0.61
4C	1.29	0.57	1.26	−0.14	2.03	0.73	1.09	0.71	1.59	1.86	2.19	0.73	0.18	0.81
5A	1.73	0.96	0.91	0.03	2.71	1.96	1.78	0.34	2.55	0.40	2.89	−0.69	0.16	0.45
5C	2.29	−0.21	1.11	−0.92	2.64	1.25	2.15	1.12	2.53	0.52	1.97	−0.67	0.20	−0.23
6A	1.08	0.74	1.92	−0.07	2.40	0.40	1.66	−0.29	1.88	0.72	2.50	−0.68	1.57	−0.20
6C	1.69	0.21	1.76	−0.25	2.72	1.56	2.17	0.79	2.58	0.24	2.06	−1.24	0.69	0.62

The first two bias columns (highlighted in orange) represent the medium scenario (N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.8, N^I=2N^A). The cells highlighted in red show increases in the absolute bias, while cells highlighted in yellow show decreases in absolute bias, compared with the medium scenario design. The other scenarios maintain the same factors as the medium scenario design except for the area and interviewer variances which are specified above

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 across factor changes. 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 sample sizes and equal number of interviewers and areas. These results are included in Appendix D.

To summarise, the main points on the percentage relative bias are the following:

- In most scenarios N=5760, the relative percentage bias is around 1-2% once interviewers are allocated work in at least two areas.
- High biases are obtained when interviewers work in solely in one area. Biases are reduced once interviewers are allocated to at least two areas.
- The area random parameter bias is almost always greater than the interviewer random parameter bias, highlighting the influence of group size on estimator accuracy.
- Greater biases are observed for smaller sample sizes.
- Scenarios with equal numbers of areas and interviewers generally obtain higher biases for both variance parameter estimators than scenarios with twice the number of interviewers to areas. A higher number of clusters, as well as an unequal number of clusters for the two classifications, can explain the more accurate result for the latter scenarios.
- 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.

III.5.4. 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 as well as very few cases of over-coverage, especially for scenarios

where each interviewer works only in one area. Under-coverage is observed for most $N^I=2N^A$ scenarios when the interviewer is working in only one area (Table 12, Row 1). In these cases the under-coverage is generally more pronounced for the σ_v^2 than the σ_u^2. This highlights the positive effect of a high number of higher-level units on the confidence interval coverage. The lowest coverage rates of 87% and 88% are obtained for the following scenarios respectively: N=5760, $N^I=2N^A$, σ_v^2=0.2, σ_u^2=0.3, π=0.8, CASE 1 and N=2880 or N=1440, $N^I=2N^A$, σ_v^2=0.3, σ_u^2=0.3, π=0.8, CASE 1, for the σ_v^2.

Table III.12: Wald 95% Confidence Interval Coverage by Sample Size and Interviewer Allocation for $N^I=2N^A$ Scenarios

	Sample Size					
	5760		2880		1440	
IA	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2
1	91.4	93.8	90.1	93.6	87.7	91.0
2A	94.5	95.0	92.9	93.5	91.2	91.1
2B	96.0	92.4	94.0	94.1	92.9	91.8
2C	95.1	94.1	93.3	92.8	92.6	91.0
3A	93.8	94.7	92.8	94.3	93.7	92.5
3B	95.0	94.0	94.1	94.3	92.7	92.7
3C	94.6	93.4	94.1	93.8	92.8	89.9
3D	95.9	93.4	93.0	94.1	92.7	91.5
3E	94.6	95.0	93.7	94.4	92.4	91.1
3F	94.8	93.6	95.0	95.6	93.3	92.0
3H	93.9	94.0	94.1	93.1	93.4	91.2
4A	95.2	94.5	94.5	93.0	92.9	91.2
4B	94.4	95.6	94.0	93.5	92.3	91.3
4C	94.1	95.0	94.5	95.5	92.7	92.6
5A	95.2	94.8	94.8	93.6	94.1	92.7
5B	95.2	94.7	94.3	93.8	93.1	93.6
5C	95.5	94.8	94.1	94.9	92.9	91.8
6A	95.1	95.1	93.6	94.6	93.5	91.5
6B	96.0	93.9	93.9	94.0	93.5	92.0
6C	94.9	95.2	94.9	94.5	93.7	92.5

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

Very high over-coverage, of approximately 100%, is present in all scenarios where $N^I=N^A$ (N=5760 or 2880 or 1440, σ_v^2=0.3, σ_u^2=0.3, π=0.8) and

where the interviewer is working in one area for both σ_v^2 and σ_u^2 (Table 13, Row 1). Over-coverage only occurs for the first interviewer case allocation scheme, as for the sparser interviewer case allocation schemes under-coverage of confidence interval for the random parameters is observed more frequently (across the various interviewer case allocation schemes) for smaller N.

Slightly lower coverage rates are observed for smaller N in most scenarios for both σ_v^2 and σ_u^2 (Table 12 & 13). The differences between the 5760 sample size scenario and the 2880 sample size scenarios are not substantial, and occasionally slightly higher coverage rates can be observed for the 2880 sample size scenarios for σ_u^2, especially for the $N^I=N^A$ scenarios. On the other hand, the 1440 sample size scenarios always obtain lower coverage rates than the other two larger sample size scenarios. These results indicate that 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.

Table III.13: Wald 95% Confidence Interval Coverage by Sample Size and Interviewer Allocation for $N^I=N^A$ Scenarios

	Sample Size					
	5760		2880		1440	
IA	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2
1	99.7	99.7	100	99.9	99.7	99.7
2	96.0	93.4	93.3	94.3	92.0	91.3
3	94.7	93.7	94.9	94.4	92.5	89.2
4	95.4	94.0	93.5	94.3	94.1	93.1
5	95.4	94.2	94.2	95.2	92.5	93.0
6	95.2	94.2	93.1	94.4	93.6	91.8

Constant factor values: σ_v^2=0.3, σ_u^2=0.3, π=0.8, $N^I=N^A$
$N^I=120$ 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

In comparing the results of Tables 12 and 13, 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 (Table 14) or by the higher-level variances. Neither do the results show any evidence of extent of interviewer overlap influencing coverage rates (Table 12 & 14). Unexpectedly, the results do not provide any evidence that the MCMC credible quantiles perform consistently better than the intervals based on asymptotic normality. Consequently, the MCMC quantiles are not presented in the results section, but are included in Appendix E. 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 confidence interval coverage for the MCMC credible quantiles in comparison to the Wald confidence interval may have been observed.

Table III.14: Wald 95% Confidence Interval Coverage by Overall Probability and Interviewer Allocation

	Sample Size					
	0.7		0.8		0.9	
IA	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2
1	93.8	95.4	91.4	93.8	91.5	93.7
2A	95.1	93.6	94.5	95.0	94.2	93.1
2C	94.6	95.2	95.1	94.1	94.9	93.0
3A	97.5	95.7	93.8	94.7	94.3	93.2
3E	94.3	93.9	94.6	95.0	94.6	94.3
3H	95.1	95.1	93.9	94.0	94.9	95.0
4A	95.4	94.8	95.2	94.5	94.1	93.7
4C	96.2	94.8	94.1	95.0	95.0	94.7
5A	95.9	94.7	95.2	94.8	94.7	93.9
5C	94.8	94.8	95.5	94.8	94.9	94.5
6A	95.8	94.5	95.1	95.1	94.2	92.8
6C	95.2	94.8	94.9	95.2	93.7	95.0

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

To summarise, the main points on the Wald confidence interval coverage are the following:

- The confidence interval coverage is close to the nominal 95% for most scenarios.
- Some cases of under-coverage (lowest observed rate is 87%) and over-coverage (highest observed rate is 100%) are obtained for scenarios allocating interviewers to just one area.
- Lower coverage rates across are observed for the scenario with the smallest sample size of 1440 cases.
- For the interviewer parameter better coverage rates are obtained for the scenarios including double the number of interviewers to areas in comparison to scenarios including equal numbers of interviewers and areas. This result highlights the better coverage properties with a larger number of clusters.
- There is no consistent pattern in the coverage rates by the overall probability, the higher-level variances or the extent of interviewer overlap.

III.5.5. Standard Errors

The precision of both fixed effect and random effects estimators is affected by N (Table 15). 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.

Standard errors for the variance estimators decrease with greater interviewer dispersion up to a certain number of areas per interviewer (Table 15). This threshold varies by N and the ratio of interviewers to areas. For the $N^I=2N^A$ scenarios for N=5760, decreases in standard errors of at least 0.05

magnitude for greater interviewer dispersion are only present for up to CASE 2, while for the 2880 and 1440 sample size scenarios a decrease is noticeable up to CASE 3. The 5760 and 2880 sample size scenarios with $N^I = N^A$ show decreases in standard errors of at least 0.05 in magnitude up to CASE 4, and up to CASE 5 for the 1440 sample size scenario. The standard errors for the interviewer variance estimators are consistently higher for the $N^I=N^A$ scenarios compared to the $N^I=2N^A$ scenarios for a specific N and interviewer allocation. This difference is expected since in the $N^I=N^A$ scenarios there are only 120 interviewers, compared to the 240 interviewers included in the $N^I=2N^A$ scenarios. 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.

Table III.15: Standard Errors by Sample Size, Interviewer Allocation and Ratio of Interviewers to Areas

	$N^I=2N^A$ scenarios								
	Sample Size								
	5760			**2880**			**1440**		
	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$
1	0.071	0.094	0.085	0.111	0.148	0.143	0.143	0.191	0.184
2	0.072	0.070	0.063	0.102	0.104	0.094	0.144	0.153	0.134
3	0.072	0.067	0.060	0.102	0.097	0.087	0.144	0.140	0.123
4	0.072	0.065	0.059	0.102	0.095	0.085	0.144	0.143	0.126
5	0.072	0.064	0.058	0.102	0.093	0.084	0.145	0.143	0.125
6	0.072	0.064	0.059	0.102	0.092	0.084	0.145	0.142	0.123
	$N^I=2N^A$ scenarios								
1	0.080	0.252	0.252	0.114	0.273	0.273	0.165	0.318	0.317
2	0.080	0.077	0.076	0.114	0.111	0.112	0.164	0.171	0.169
3	0.081	0.073	0.075	0.116	0.107	0.112	0.165	0.165	0.167
4	0.080	0.067	0.067	0.113	0.096	0.098	0.164	0.153	0.153
5	0.080	0.065	0.065	0.114	0.095	0.096	0.163	0.147	0.147
6	0.080	0.064	0.064	0.114	0.095	0.094	0.163	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

Despite N^A always being held constant the standard errors for the σ_v^2 estimator are only equal for comparable $N^I=N^A$ and $N^I=2N^A$ scenarios when the interviewers are working in at least four areas for the 5760 and the 2880 same size scenarios, and five areas for the 1440 sample size scenario. For more restrictive interviewer allocations, higher standard errors for the σ_v^2 estimator are obtained for $N^I=N^A$ scenarios compared to $N^I=2N^A$ scenarios. Again, this result highlights the possibility that an unequal number of higher-level units for the two higher-level classifications yields better estimates when interpenetration is limited.

The intercept estimator standard error is constant for all the interviewer allocation schemes for constant values of both N^I and N^A. However, for the $N^I=N^A$ scenarios as compared to the $N^I=2N^A$ scenarios for the same N and for the same interviewer case allocation scheme, greater standard errors are observed (e.g. 0.08 compared to 0.072 for the 5760 sample size scenario). This difference in standard errors is greater for smaller sample size scenarios, increasing from an average of 0.008 for the 5760 sample size to 0.02 for the 1440 sample size scenario. This result can also be explained in terms of lower standard errors for scenarios with unequal numbers of higher-level units.

Table 16 shows no evidence of any effect of the extent of interviewer overlap on the standard errors for either the intercept or the variance estimators. On the other hand, results in Table 16 show higher standard errors for all three parameters for higher overall probabilities, with some increase from $\pi=0.7$ to $\pi=0.8$, and a much higher increase from $\pi=0.8$ to $\pi=0.9$, especially for the CASE 1 interviewer case allocation scheme. This non-linear result is similar to the effect of overall probability on $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$, which shows that a greater increase in the correlation 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.

Table III.16: Standard Errors by Overall Probability and Interviewer Allocation

	Overall Probability								
	0.7			0.8			0.9		
IA	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$	$\widehat{\beta_0}$	$\widehat{\sigma_v^2}$	$\widehat{\sigma_u^2}$
1	0.069	0.087	0.077	0.071	0.094	0.085	0.079	0.112	0.112
2A	0.069	0.065	0.057	0.072	0.070	0.063	0.080	0.084	0.081
2C	0.069	0.064	0.056	0.072	0.069	0.062	0.080	0.082	0.081
3A	0.069	0.062	0.055	0.072	0.067	0.060	0.080	0.079	0.077
3E	0.069	0.061	0.054	0.072	0.066	0.060	0.080	0.079	0.077
3H	0.069	0.061	0.054	0.072	0.065	0.059	0.080	0.078	0.077
4A	0.069	0.060	0.054	0.072	0.065	0.059	0.080	0.077	0.075
4C	0.069	0.060	0.053	0.072	0.064	0.059	0.080	0.077	0.075
5A	0.069	0.060	0.053	0.072	0.064	0.058	0.080	0.076	0.075
5C	0.069	0.059	0.053	0.072	0.064	0.058	0.080	0.077	0.075
6A	0.069	0.058	0.053	0.072	0.064	0.059	0.080	0.076	0.074
6C	0.069	0.059	0.053	0.072	0.063	0.058	0.080	0.075	0.074

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 17 shows that when the value of the variance changes the standard error changes in the same direction for the respective variance estimator as well as the intercept estimator. The unchanged variance does not experience changes in the estimator standard errors, once interviewers work in at least two areas. For example, when comparing scenario H (σ_v^2=0.3, σ_u^2=0.4) to the medium scenario design A (σ_v^2=0.3, σ_u^2=0.3), a slight increase in the standard errors can be noticed for the intercept estimator while a substantial increase can be observed in the standard errors of the interviewer variance estimator. On the other hand, for the area variance estimator an increase is only noticeable for CASE 1.

Table III.17: Standard Errors by Different Scenarios

IA	A: σ_v^2=0.3, σ_u^2=0.3	F: σ_v^2=0.2, σ_u^2=0.2	G: σ_v^2=0.4, σ_u^2=0.4	H: σ_v^2=0.3, σ_u^2=0.4	I: σ_v^2=0.4, σ_u^2=0.3	J: σ_v^2=0.3, σ_u^2=0.2	K: σ_v^2=0.2, σ_u^2=0.3
				$\widehat{\beta_0}$			
1	0.071	0.061	0.080	0.074	0.077	0.068	0.064
2A	0.072	0.062	0.081	0.075	0.078	0.069	0.065
3A	0.072	0.062	0.081	0.075	0.078	0.068	0.065
4A	0.072	0.062	0.081	0.075	0.078	0.068	0.065
5A	0.072	0.062	0.081	0.075	0.078	0.069	0.065
6A	0.072	0.062	0.081	0.075	0.078	0.069	0.065
				$\widehat{\sigma_v^2}$			
1	0.094	0.071	0.118	0.105	0.106	0.077	0.079
2A	0.070	0.052	0.088	0.072	0.085	0.067	0.054
3A	0.067	0.050	0.082	0.067	0.081	0.065	0.051
4A	0.065	0.049	0.081	0.066	0.080	0.064	0.049
5A	0.064	0.048	0.080	0.064	0.079	0.063	0.048
6A	0.064	0.048	0.079	0.064	0.078	0.063	0.048
				$\widehat{\sigma_u^2}$			
1	0.085	0.070	0.101	0.103	0.083	0.062	0.085
2A	0.063	0.051	0.075	0.075	0.064	0.051	0.062
3A	0.060	0.049	0.072	0.071	0.061	0.049	0.059
4A	0.059	0.048	0.071	0.070	0.059	0.048	0.058
5A	0.058	0.047	0.070	0.069	0.059	0.047	0.058
6A	0.059	0.047	0.069	0.068	0.058	0.047	0.057

Scenario A represents the medium scenario design (N=5760, N^I=240, N^A=120, σ_v^2=0.3, σ_u^2=0.3, π=0.8, N^I=2N^A). The other scenarios maintain the same factors as the medium scenario design except for the σ_v^2 and σ_u^2 which are specified above

To summarise, the main points on the standard errors are the following:

- Standard errors increase for both the intercept and the variance parameter estimators when the total sample size is decreased.
- When the number of areas and interviewers are equal the standard errors for the area and interviewer variance estimators are the same when the variances are equal. For scenarios with twice the number of interviewers to areas, the area parameter estimator has larger standard errors than the interviewer parameter estimator, when the two variances

are equal. This result confirms the negative relationship between the number of higher-level units and the precision of the estimator.

- The standard errors of the variance estimators decrease with greater interviewer dispersion, up to a threshold number of areas per interviewer, which varies by sample size and the ratio of interviewers to areas.
- Higher standard errors are obtained for scenarios with equal numbers of areas and interviewers compared to scenarios with double the number of interviewers to areas. This result highlights the increased precision for scenarios with unequal number of higher-level units for the two higher-level classifications.
- 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 an increase in the value of the variance the standard errors also increase for the respective variance estimator as well as the intercept estimator.

III.5.6. Extreme Case Allocations

This paper only considers interviewer allocation schemes which restrict interviewers to six areas or less. This restriction makes pragmatic sense. However, just for analytical purposes, for the 2880 sample size and 1440 sample size scenarios extreme case allocation schemes are considered. For this extreme allocation scheme each case for each interviewer is situated in a different area. The scenarios with smaller N have been chosen since sample size and the number of higher-level units has been shown to have the greatest influence on the quality of the estimators. Therefore, any improvements in the estimators with further interviewer dispersion should be noticeable for these scenarios.

Table III.18: Properties of the Estimators and Test Statistic by Scenario and Interviewer Allocation

	Bias of Estimator		Coverage of CI		Power of Wald Test		Standard Errors		Correlation
	$N^I=2N^A$ Scenarios								
	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2	σ_v^2	σ_u^2	
IA	1440 sample size scenario								
6A	7.42	2.42	93.5	91.5	0.92	0.88	0.142	0.123	0.034
6B	6.04	0.97	93.5	92.0	0.91	0.90	0.140	0.122	0.040
6C	6.12	1.09	93.7	92.5	0.91	0.89	0.140	0.123	0.043
EXTREME	5.88	1.25	91.8	92.1	0.92	0.91	0.137	0.120	0.063
	2880 sample size scenario								
6A	3.05	1.78	93.6	94.6	1	1	0.092	0.084	0.034
6B	2.99	1.36	93.9	94	1	1	0.092	0.084	0.041
6C	4.27	0.89	94.9	94.5	1	1	0.093	0.083	0.045
EXTREME	2.85	1.20	94.0	93.8	1	1	0.090	0.082	0.067
	$N^I=N^A$ Scenarios								
	1440 sample size scenario								
6	10.34	6.73	93.6	91.8	0.91	0.88	0.147	0.144	0.007
EXTREME	9.07	8.24	94.0	93.7	0.94	0.94	0.139	0.138	0.053
	2880 sample size scenario								
6	3.8	2.99	93.1	94.4	1	1	0.095	0.094	0.002
EXTREME	3.58	3.93	94.1	94.2	1	1	0.090	0.090	0.052

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

In Table 18, the quality of the estimators for the extreme case can be compared with those for CASE6. The results do not indicate any dramatic improvements in the quality of estimators for extreme interviewer case allocation schemes. This observation confirms previous results discussed in this paper, which suggest that beyond a limited number of areas to interviewers, further dispersion does not yield any gains in the quality of the estimators. For the simulated data considered, for large sample sizes of around 6000 cases, this threshold is equal to three areas per interviewer.

III.6. Discussion

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 drastically 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, very 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 $N^I = 2\ N^A$ 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 $N^I=2N^A$ 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 the 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. The overall probability of the outcome variable is not something the data analyst can easily change through the survey design, unlike other factors such as N^I which are more easily changed. Therefore any results indicating a negative effect of higher prevalence of the outcome variable on the quality of the random parameter estimates are more problematic. However, 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. On the other hand, interviewer dispersion does not mediate the positive effect of higher overall probability on the standard errors of estimators. 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 $corr\left(\widehat{\sigma_u^2}, \widehat{\sigma_v^2}\right)$, although this is no longer present for higher-level effects with a large number of clusters in the sample. In fact, lower negative $corr\left(\widehat{\sigma_u^2}, \widehat{\sigma_v^2}\right)$ are observed for higher area variances σ_v^2 up until interviewer allocation CASE 3A, but no consistent change is observed for higher interviewer variances σ_u^2 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 – which are not under the control of the researcher – 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 medium 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 $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ 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. The results for the extreme case allocations do not show any dramatic improvements in the properties compared with CASE 6, confirming that beyond a certain number of areas to interviewers, very often being equal to three areas per interviewer for large N, further dispersion does not yield any gains in the properties.

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 $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$ values for 2 areas per interviewer allocation schemes, with less overlap resulting in lower negative $corr\left(\widehat{\sigma_u^2},\widehat{\sigma_v^2}\right)$. 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.

III.7. Conclusion

For all the properties considered a trend can be identified – for interviewer case allocation schemes beyond a certain number of areas each interviewer, generally 3 areas per interviewer for a medium or large sample size (2880 or 5760 cases), the effect of other factors on the properties observed for more geographically-restrictive interviewer case allocation schemes cancels out. Therefore, interviewer dispersion acts as a mediating factor on the effect of interviewer overlap, the overall sample size N, the ratio of interviewers to areas, the overall probability π and the area variance σ_v^2 on the properties of the estimators and Wald test.

IV. The Effect of Sample Size and Interviewer Allocation Profiles in Longitudinal Samples on Inference from Multiple Membership Multilevel Logistic Regression Models (Paper 3)

IV.1. Introduction

Across the waves of a longitudinal survey several interviewers may contact sample members to participate in a survey. A modelling problem particular to this kind of data pertains to the inclusion of higher-level random effects for the various wave interviewers, whilst a substantive problem is the assessment of the relative importance of previous and current wave interviewers on current wave nonresponse. If all distinct interviewers from both the current and previous waves associated with a case influence the current wave response decision, failing to account for the multiple membership structure will lead to an underestimation of the interviewer variances (Goldstein, 2011).

One approach to correctly model this data and estimate the interviewer variance and to identify the relative effect of various wave interviewers for cases experiencing interviewer change on current wave nonresponse is to use multiple membership (MM) models (Lynn et al., 2013). Paradata, that includes the identification codes of the interviewers allocated to each case at each wave, is required. Such models allow the effect of all distinct interviewers associated with a case to be incorporated in the model by attributing a weight to each interviewer effect. These weights represent each interviewer's relative effect. The choice of weights is either based on theory or an empirical assessment using the Deviance Information Criterion (DIC), as proposed in Goldstein (2011) and advocated in Lynn et al. (2013). The DIC is used to select the best fitting model among different competing possibly non-nested Bayesian models, with higher DIC values indicating a poorer model fit (Spiegelhalter et al., 2002).

Models including various weight specifications are fitted and the weights corresponding to the model with the lowest DIC are retained.

Fitting these models with various weights to data from the UK Family and Children Study (Lyon et al., 2007) in the first paper showed no improvement in comparison to a simple multilevel logistic model that only included an effect for the current wave interviewer. This result may indicate that it is only the current wave interviewer who has an impact on current wave cooperation. The current interviewer's direct interaction with and influence on the immediate experience may cancel out any previous wave experience. Alternatively, this result may be due to a lack of power to detect the MM structure, either because insufficient cases experienced interviewer change, or because the higher-level variance is small, or a combination of both factors. Questions regarding the reliability of the DIC measure for determining the true MM weights and the properties of estimators and the power of significance tests for MM models with different weight specifications have not been explored yet.

IV.2. Study Aims

This study, through simulation studies, assesses the reliability of the DIC to detect the correct MM weights. A second aim of the study is to investigate the properties of estimators and test statistics for MM models with different weight specifications across a range of scenarios. The reliability of the DIC will be assessed in terms of the percentage of times the models with the correct model weights correspond to the lowest DIC value. The properties considered include the percentage relative bias, the standard error, the confidence interval coverage and the power of significance tests. The properties of the MM models are investigated when weights are chosen on an a priori theoretical basis and alternatively when weights are chosen on the basis of the DIC. The different scenarios considered will vary in terms of the true MM weights (from unequal weights of 0.9 and 0.1 to equal weights of 0.5 each), the different interviewer

change profiles and the proportions of cases experiencing interview change. Different total sample sizes and number of interviewers (groups) are also considered. Other factors, such as the overall probability and the higher-level variances will be held constant at a realistic value.

The study will attempt to identify in which scenarios choosing the weights based on the lowest DIC produces acceptable properties of estimators and test statistics and the correct weights selection. These properties will be compared to those for equivalent models including weights based on theory. Models with theoretically-based weights will include both correct weights and incorrectly specified weights (with varying degrees of misspecification being considered). Interviewer allocations for the two waves considered (the current wave at which nonresponse is being analysed and the previous wave) aim to represent some of the possible extreme interviewer work allocations and change profiles. Starting values for the other factors are meant to represent typical survey values. These scenario specifications aim to make the results relevant to survey practice.

Although the factor conditions and the application considered are very specific and restricted to survey design and the exploration of interviewer effects on nonresponse, the same MM structure and the question of how best to choose the model weights may arise in other settings. For example, a study may wish to explore the influence of a pupil's secondary school on the pupil's probability to go on to further education. Pupils who have attended more than one school during their secondary years of schooling have a MM structure, and the relative effect of the final and previous school can be assessed using MM models. The results from this study would have implications for league tables and funding. Other applications may include studies of neighbourhood effects on the propensity to seek traditional birth assistants in sub-Saharan Africa, studies on the influence of religious group affiliation on the likelihood of doing volunteering work, receipt of unemployment benefits with changing household membership in longitudinal studies, clinical trials investigating the probability of rehabilitation after receiving care from a combination of practitioners, and veterinary studies considering the influence of flock memberships on disease contagion.

The simulation results from this paper on the percentage of cases with multiple memberships required for adequate estimates of the higher-level variance and the probability of the DIC measure identifying the correct weights for various data structures will highlight any inference problems arising for MM models. The performance of the DIC in choosing between competing MM model weights will suggest whether the substantive interpretation of the weights based on the DIC can be emphasised. The results will be of particular importance when using modelling not simply for explanation purposes but also for prediction. This study will also indicate under which scenarios choosing weights based on an empirical assessment method compared to relying on theory yields better estimator properties and power of Wald test. This study may also inform the design of studies with MM structures.

IV.3. Background

Model selection refers to the process of choosing the best model for the data being analysed between competing models. Model selection tools allow the ranking of different competing models, enabling model selection. The properties of parameter estimators can be sensitive to model specification, particularly to the omission or misspecification of the higher-level structure (Chung & Beretvas, 2012; Luo & Kwok, 2009; Meyers & Beretvas, 2006; Tranmer & Steel, 2001). Consequently, it is important to base model selection on an empirical assessment method or on strong theory which has been rigorously tested. When a strong theoretical basis for the model structure is lacking, model selection has to be solely based on an empirical assessment method. Consequently, the consistency with which the model selection method identifies the true model as being the correct model across different applications, data structures and types of models must be examined. Moreover, the properties of estimators when the model choice is based on this selection method also need to be investigated.

The model selection tool used in this paper is the DIC, a Bayesian model selection tool which takes into consideration both the goodness-of-fit (posterior distribution of the deviance) and the complexity of the model (effective number of parameters), which is particularly appropriate for models including hierarchical parameters estimated using Markov Chain Monte Carlo (MCMC) (Spiegelhalter et al., 2002). Some studies have analysed the performance of the DIC for different subject areas and model types, including spatial models for medical data (Zhu & Carlin, 2000), stochastic volatility models for financial time series data (Berg et al., 2004), catch-at-age models for fisheries stock assessment data (Wilberg & Bence, 2008), hierarchical threshold mixed models for genetic analysis of veterinary data (Kizilkaya & Tempelman, 2003) and discrete-time population models for ecological data (Ward, 2008). These studies generally show that the DIC measure performs well in detecting the true model or similar models which adequately represent the data. At present there is no literature on the performance of the DIC for selecting between different MM weights.

There is very limited literature which explores the estimator properties and power of a significance test for MM models with different data structures. Browne et al. (2001) look at the properties of estimators for MM models using a simple simulated education data example. For this example, students' attainment is the outcome variable. Students are associated with the schools (higher-level units) they have attended. Only 10% of a total of 3435 students have attended two schools, whereas the other 90% are only associated with one higher-level unit. The higher-level variance is set to 0.1, whilst the individual-level variance is set to 0.6. The authors find that when using MCMC estimation with diffuse priors the mean point estimate from the posterior distribution has very low bias, and the interval estimates based on the percentiles of the chains for the posterior distribution have coverage very close to the nominal 95% value. The Iterative Generalised Least Squares (IGLS) estimation results fare less well. A limitation of this study is that the bias and confidence interval coverage of the estimators in the case of incorrectly specified model weights are not considered. The authors only consider a case for true MM weights of 0.5 and 0.5 and specify model weights to be the correct weights. In real life scenarios

the true weights are not known and the model weights may therefore be subject to misspecification.

Chung and Beretvas (2012) run simulation studies to analyse the impact of incorrectly specifying a simple two-level model (which only takes account of the most recent higher-level unit) for MM data for a continuous outcome. This impact is assessed in terms of the bias for both fixed and random effects coefficients. Both 2-level models and MM models are fitted for all scenarios and the biases are compared for the two model specifications. The values for the following factors are varied: the intra-class correlation, the number of higher-level units, the cluster size, the percentage of cases with multiple memberships (change cases) and the number of multiple memberships for the change cases. Where a substantial bias is observed, for either the 2-level model or the multiple membership model, an ANOVA test is performed to identify which of the above-mentioned factors is significantly associated with the observed bias.

When a MM model is specified no bias is observed for any of the parameters considered. No substantial difference in the relative percentage bias for the overall intercept and a level-one fixed effect was observed between the perfectly nested model and the MM model specification. Biases are observed for the level-two fixed effects and the two random effects when a purely nested model is run. A significant negative bias for the level-two fixed effect was observed for the 2-level model specification ignoring the MM structure of the data. This bias was higher for a higher proportion of cases being associated with multiple memberships (higher percentage of students who changed schools) as well as a higher number of multiple memberships (mobile children attending three rather than two different schools). These two factors (degree of mobility across student population and number of schools attended by mobile students) interact with each other, such that with high values for both factors a more substantial negative bias is obtained for the level-two fixed effect. The individual-level variance is overestimated when a perfectly hierarchical model is specified. A higher percentage of cases with multiple memberships and a larger number of multiple memberships for the change cases are found to be associated with larger positive bias. The higher-

level variance is underestimated when the multilevel structure is wrongly specified, with a higher percentage of cases experiencing change being significantly associated with a larger negative bias.

The underestimation of the higher-level variance when the MM structure of the data is ignored has been documented by Goldstein (2011, Chapter 13). Chung and Beretvas (2012) argue that the variance not captured in the higher-level model has to be apportioned to the individual-level variance, which is overestimated to allow for the correct estimation of the total variance. This is in fact observed in the simulation study results and documented in other simulation studies exploring the effect of omitting a level in a multilevel model (Tranmer & Steel, 2001; van den Noortage et al., 2005).

Some studies using multiple membership models with real data to investigate substantive questions make some reference to the robustness of the parameter estimates across different weight specifications. Fielding (2002) and Fielding and Yang (2005) investigate the influence of multiple teachers or educational institutions on individual students' educational achievement. Both studies base the calculation of weights on the proportion of time spent with each higher-level unit. Both papers assert that the accuracy of parameter estimates of MM models is not sensitive to the model weights specifications. Similarly, Goldstein (2011b), in analysing the influence of multiple applicants on the grades awarded to research grant applications, finds that the results are stable across different weighting schemes. These studies do not give any detail as to the weighting profiles attempted and the estimates obtained. Consequently, the reported stability across weighting profiles probably reflects attempted weighting profiles which are close to the correct weights. As the model weights specified deviate from the correct weights, and the sum of the square of these model weights deviates from this measure for the correct weights, the estimated variance will surely be biased.

Chung and Beretvas (2012) highlight the need for further simulation studies investigating the properties of estimators for misspecified MM models. This current study addresses this lacuna in the literature, and also investigates

the performance of the DIC for detecting the correct MM weights – a research topic which has not been explored yet.

IV.4. Methodology

This section presents the details of the simulation design. The first section presents the MM multilevel logistic regression model being fitted to the simulated data. In the next section, the process by which data is generated is explained in detail. The various simulation scenarios and the design factor values considered are then specified. Next, the stored quantities from each fitted model are listed. The formulas used to calculate the properties of the estimator and test statistic, and the reliability of the DIC from these stored quantities are then presented. The various models – models including various different theoretical weights and the models including weights based on the DIC – for which the above-mentioned measures are calculated are also specified in this section.

IV.4.1. Simulation Model

The following model is used:

$$\text{logit}\left(p_{ij_c j_p}\right) = \beta_0 + w_{ij_p} u_{j_p} + w_{ij_c} u_{j_c}, \qquad w_{ij_p} + w_{ij_c} = 1 \qquad \text{(IV. 1)}$$

where $p_{ij_c j_p}$ is the probability of individual *i* interviewed by interviewer j_p at the previous wave and interviewer j_c at the current wave refusing to participate at wave *n* and the interviewer-specific residuals u_j for both the current and previous wave interviewers come from one distribution N(0, σ_u^2). Cases experiencing interviewer change have a weighted average effect of the previous and current wave interviewer effects. The model weights for the current and previous wave interviewers are represented by the terms w_{ij_c} and w_{ij_p} respectively. Cases with the same change profile are given the same MM

weights. Consequently there will only be two possible pair of weights for each scenario and model weight profiles – one for the interviewer continuity cases and one for the interviewer change cases. While cases allocated to the same interviewer across both waves are given a weight of 1 for w_{ij_p} and a weight of 0 for w_{ij_c}, cases experiencing an interviewer change have two non-zero weights (identical for all cases) summing to 1. The terms w_{ij} and W_{ij} will be used to refer specifically to the pair of model weights and the pair of true MM weights for change cases. In total ten model weight profiles will be considered for each scenario. Nine of these profiles represent different possible theoretically-based weighting schemes. The other weighting profile is based on the DIC.

For each scenario nine weight profiles are specified, and consequently nine models are fitted using each simulated dataset. Each model will include different model weights for the cases experiencing interviewer change, w_{ij}. These weight profiles vary by 0.1, from weights of (0.9, 0.1) to (0.1, 0.9). For one of these nine weight profiles the model weights w_{ij} are the correct weights, equal to the true MM weights W_{ij} (the weights used to generate the data), while the other eight models will have incorrect w_{ij}, with varying degrees of misspecification. These nine weight profiles represent the different possible theoretically-based weights. After all 9 models are fitted, the model corresponding with the lowest DIC is chosen. This is repeated for all 1000 simulations for each scenario. The 1000 models (from a total of 9000 models) with the lowest DIC will include different weighting profiles. Their one common criterion is that they provide the best fit (determined by the DIC value) for each particular simulated dataset.

STATA Version 12 calling MLwiN Version 2.25 through the 'runmlwin' command (Leckie & Charlton, 2011) is the software used to fit the models to the simulated data. Models are fitted using the MCMC estimation method with diffuse priors, a burn-in length of 5,000 and 100,000 iterations. The burn-in length and number of iterations were chosen by running a sensitivity analysis for a few scenarios prior to starting the main analysis. The second order penalised quasi-likelihood (PQL) estimates provide initial values for parameters. Due to the computational power and efficiency requirements of MCMC

estimation, and the large number of models estimated, the IRIDIS High Performance Computing Facility, and the associated support services at the University of Southampton have been required to complete the work.

IV.4.2. Data Generating Procedure

Since the study is mainly concerned with the properties of the estimator for the interviewer random parameter only an overall intercept β_0 is included as a fixed effect. The regression coefficient for this overall intercept β_0 is determined after considering the overall probability of the outcome for the mean interviewer membership, π, and substituting it in the following formula:

$$\beta_0 = \log_e \frac{\pi}{1-\pi}. \qquad (\text{IV}.2)$$

This value is constant across all cases. Then an interviewer random effect is generated from a normal distribution of mean 0 and variance σ_u^2 for each interviewer included in the analysis. If for example the previous wave includes 100 distinct interviewers and the current wave includes another 20 distinct interviewers not present in the previous wave, a total of 120 interviewer effects are generated. The true MM weights W_{ij_p} and W_{ij_c} are specified. As explained above, cases with no interviewer change will be allocated (1, 0) weights, whilst cases with interviewer changes are allocated two non-zero weights (W_{ij}) which sum to unity. These non-zero weights are maintained constant across all change cases. The log-odds of each case, η_{ij}, are computed by adding the overall intercept value to the weighted average of the simulated random effects:

$$\eta_{ij} = \beta_0 + W_{ij_p} u_{j_p} + W_{ij_c} u_{j_c}. \qquad (\text{IV}.3)$$

These values are then converted to probabilities using the equation:

$$p_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}. \qquad (\text{IV}.4)$$

Values of the dependent variable Y_{ij} for each case, are generated from a Bernoulli distribution with probability p_{ij}. The dependent variable is a dichotomous one, with 0 representing nonresponse and 1 representing response to the survey request. For each scenario 1000 simulations are generated using R Version 2.11.1. For each simulated dataset the simulation model presented in the previous section is run nine separate times, each time specifying different w_{ij}.

The procedure used for generating the data for one specific scenario is presented in Appendix F. The same programming code, with the appropriate changes to the factor values and the change in the code for the interviewer allocations, can be used for other scenarios. Appendix G presents the different codes for obtaining the interviewer codes under different change profile types.

IV.4.3. Simulation Scenarios and Factors

The scenarios considered include the following factors: the overall sample size N, the number of interviewers at the previous and current waves N_p^I and N_c^I, and by consequence the number of cases per interviewer, the percentage of cases with interviewer change (percentage change), the interviewer change profile type, the interviewer variance σ_u^2, the overall probability of the outcome variable π, and the true MM weights W_{ij_p} and W_{ij_c}. The following factor values will be considered typical values and maintained constant across the majority of scenarios: N=5760, N_p^I=240, 24 cases per interviewer at the previous wave, σ_u^2=0.3, π=0.8. While maintaining these values, the other factor values will be altered to assess the effect of different percentage change, change profiles and W_{ij} on the properties of the estimator and test statistic and the DIC reliability measure for realistic general household survey scenarios.

Six change profile types will be considered here. Their characteristics are outlined briefly in Table 1 and described in more detail in the main text. These profile types aim to represent different plausible, yet extreme, interviewer allocations, with the intention of covering various possible interviewer work allocations. It is appreciated that these scenarios do not

provide an exhaustive overview of all possible allocations. However, they give some indication of the impact of different change profiles. In the previous wave all interviewers have 24 cases allocated. This previous wave caseload is maintained across all scenarios.

In Type A and Type B scenarios the percentage change refers to the proportion of cases of each previous wave interviewer which are allocated to a different interviewer in the current wave. The Type A scenarios include the same pool of interviewers for both waves. At each wave the same interviewers are present, with the same workload, but each interviewer loses a specific amount (represented by the percentage change factor) of cases from the previous wave cases which are allocated to different interviewers in the current wave. The previous wave interviewer code for the change cases is collected and re-distributed randomly across change cases in the current wave. The Type B scenarios include a new pool of interviewers at the current wave to whom change cases are allocated. Each interviewer in the previous wave has a particular percentage of cases removed. The new pool of interviewers each have a caseload equal to the number of cases removed from each previous wave interviewer. The new interviewers introduced in the current wave are allocated the change cases randomly. For Type A scenarios the following percentage changes are considered: 8%, 21%, 33%, 50% and 92%, while for Type B scenarios the 8% and 50% changes are considered.

In Type C, D, E and F scenarios the percentage change refers to the proportion of interviewers who drop out of the survey and have all their cases allocated to other interviewers. The other interviewers maintain all their cases across both waves. Since it is interviewers that are being dropped the total caseload (24 cases times the number of dropped interviewers) must be equally divisible by the remaining interviewers or the newly recruited interviewers. Consequently, for these scenario types only the 50% change scenario will be considered.

In Type C scenarios the cases of the interviewers who drop out of the survey in the current wave are distributed randomly among all the other interviewers present in the previous wave. The previous wave interviewer code

for the interviewer continuity cases (no change) is collected and distributed randomly across change cases in the current wave. Consequently, the retained interviewers will double their case load in the current wave. On the other hand, in Type D scenarios newly recruited interviewers are allocated these change cases randomly in the current wave. In this case all interviewers have a caseload equivalent to the previous wave caseload, since the retained interviewers are supplemented by a group of new interviewers matching in number to the group of dropped interviewers. For scenario D the interviewer codes of the new interviewers, repeated for a number equal to the caseload, are allocated randomly to the change cases.

Table IV.1: Change Profile Type Characteristics

Type	Previous Wave	Current Wave	Percentage Change	Random
A	240 interviewers, 24 cases/interviewer	240 previous interviewers, 24 cases/interviewer	cases /interviewer	case level
B	240 interviewers, 24 cases/interviewer	240 previous interviewers and 240 new interviewers, 12 cases/interviewer	cases /interviewer	case level
C	240 interviewers, 24 cases/interviewer	120 previous interviewers, 48 cases each	number of interviewers	case level
D	240 interviewers, 24 cases/interviewer	120 previous interviewers and 120 new interviewers, 24 cases each	number of interviewers	case level
E	240 interviewers, 24 cases/interviewer	120 previous interviewers, 48 cases each	number of interviewers	interviewer caseload level
F	240 interviewers, 24 cases/interviewer	120 previous interviewers and 120 new interviewers, 24 cases each	number of interviewers	interviewer caseload level

For scenarios E and F the intact caseload of a dropped interviewer is allocated randomly to another interviewer. In Type E scenarios the remaining interviewers from the previous interviewers take on this extra workload, whilst for Type F scenarios new interviewers are introduced to take on the added

workload. The number of interviewers and their respective caseload for change profile E is the same as for change profile C, while that for F is equivalent to Type D. The difference between Type C and Type D in comparison to Type E and F is that while for the former profile types the individual change cases are distributed randomly for the latter profile types interviewer-specific caseloads are distributed randomly.

Table 2 lists the change indicator and the interviewer allocations for the previous and current waves for each case across the six change profile types for a simple example scenario. This example scenario includes 8 cases allocated to 4 interviewers with 2 cases each in previous wave with 50% change. This example has been included to help understand the different interviewer change profiles. For further details Appendix G, which includes the R codes used to generate the different interviewer allocations under the different change profile types, can be consulted.

Table IV.2: Interviewer Case Allocations for the Example Scenario

Case No	Change Indicator Type A and Type B	Change Indicator Type C – Type F	Previous Interviewer Code	Current Interviewer Code for Change Profile Type:					
				A	B	C	D	E	F
1	1	0	1	4	5	1	1	1	1
2	0	0	1	1	1	1	1	1	1
3	0	1	2	2	2	4	5	4	6
4	1	1	2	3	8	1	6	4	6
5	0	1	3	3	3	1	6	1	5
6	1	1	3	2	6	4	5	1	5
7	1	0	4	1	7	4	4	4	4
8	0	0	4	4	4	4	4	4	4

Two different real weight profiles W_{ij} are considered, one giving equal weights, W_{ij}=(0.5, 0.5), and the other giving unequal weights, W_{ij}=(0.9, 0.1), to change cases. For some Type A scenarios W_{ij}=(0.7, 0.3) is also included. In this study, the unequal weights, W_{ij}=(0.1, 0.9), will not be considered. It is believed that the trends in the estimator and test statistics properties for W_{ij}=(0.9, 0.1) and W_{ij}=(0.1, 0.9) scenarios across changes in the other factors

are similar. Moreover, on a conceptual level the previous and current wave interviewer allocations are interchangeable.

IV.4.4.Properties of the Estimator and Test Statistic and DIC Reliability Measure

The stored quantities for each model include the parameter estimates, the standard errors, the 95% confidence intervals based on the asymptotic normal distribution and the credible 95% confidence interval based on the MCMC quantiles, the Brooks-Draper diagnostic, the lower and upper bound of the Raftery-Lewis diagnostic, as well as the DIC. The properties of the estimator and test statistic and the DIC reliability measure are calculated using the data from these stored quantities. The properties include the percentage relative bias, the standard error, the confidence interval coverage and the power of the Wald test.

The DIC reliability measure is calculated as follows. For each scenario, 1000 simulated datasets are generated. For each of these 1000 datasets 9 models are fitted, each specifying different w_{ij} based on theory. For each simulation run, out of these nine models the model corresponding with the lowest DIC is selected. From a total of 9000 models run for each scenario the 1000 selected models will have different w_{ij}. The distribution of the w_{ij} for these chosen models is presented. The proportion of times the model with the correct model weights (w_{ij}=W_{ij}) is selected represents the reliability of the DIC measure for selecting the correct model weights. A less strict reliability measure quantifies the percentage of times the correct model weights or the adjacent model weights are selected.

The accuracy of a parameter estimator can be assessed by calculating the percentage relative bias using the formula

$$\frac{1}{1000}\sum_{i=1}^{1000}\frac{\hat{\theta}_i-\theta}{\theta}*100 \qquad \text{(IV. 5)}$$

where $\hat{\theta}_i$ is the parameter estimate, θ is the true parameter value and i is the simulation number. Since some variation from the true parameter value is expected due to Monte Carlo error only bias values with a minimum absolute value of 3% will be considered to be truly identifying bias. The confidence interval coverage (Maas & Hox, 2005) rate is calculated as the number of simulations for which the true parameter value lies within the 95% Wald confidence interval. The coverage rate is compared with the expected 95% rate. The precision of an estimator is assessed by calculating the mean standard error using the formula

$$\frac{1}{1000}\sum_{i=1}^{1000} S.E._i(\hat{\theta}). \qquad \text{(IV. 6)}$$

The power of a test indicates the probability that the null hypothesis is correctly rejected. Here the Wald test is used to test the null hypothesis, specifying the true parameter value to be zero. This consists of dividing the coefficient estimate by its standard error and squaring that value. The corresponding p-value for this value is obtained from a chi-squared distribution. If the p-value is greater than 0.05 then the null hypothesis is not rejected. The proportion of datasets for which the null hypothesis is retained is subtracted from 1 to obtain the power of a test.

These properties are estimated ten times – nine of which correspond to the models with w_{ij} based on theory and the other corresponding to the model with w_{ij} based on the DIC. For each scenario, the values of these measures for model with the correct weight profile (when w_{ij_c} and w_{ij_p} correspond to W_{ij_c} and W_{ij_p}) is compared to the models with the other eight incorrect models with w_{ij} based on theory as well as the model with weights based on the DIC.

The procedure for running the models, storing the output quantities and calculating the properties is specified in Appendix H. The same programming code, with the appropriate changes to the factor values and file names, can be used for other scenarios.

IV.5. Results

The sections below will present the results for the properties of the estimator and the power of the Wald test for the random effects parameter and the DIC reliability measure across various simulation scenarios. The properties for the model specifying the correct weights, that is w_{ij}=W_{ij}, will be highlighted in the tables below. Any trends in the properties of the estimator, the power of the Wald test and the DIC reliability across variations in the interviewer allocations and any other factors will be outlined. As a general note the reader is informed that the results show a lot of interactions between the various factors on the various properties considered. It is acknowledged that the detailed results presented below are specific to the particular combinations of factor values specified. For a summary of general patterns the reader is referred to the concluding bullet points at the end of each section.

IV.5.1. Percentage Relative Bias

Negligible or low relative percentage bias (of less than 4%) is observed for models specifying the correct w_{ij} across the different scenarios considered, in agreement with the result in Browne et al. (2001). As expected, models specifying incorrect w_{ij} are subject to bias. Model weights misspecification has greater negative consequences for the percentage relative bias of the variance estimator for scenarios with a higher proportion of interviewer change. This can be observed in Table 3 which presents the percentage relative bias for scenarios varying in terms of the percentage change, while holding constant these factor values: Type A, N=5760, N_p^I=240, 24 cases per interviewer at both the current and previous wave, W_{ij}=(0.5, 0.5), σ_u^2=0.3 and π=0.8. The most imprecise weighting profiles, so either w_{ij}=(0.9, 0.1) or w_{ij}=(0.1, 0.9), give rise to an underestimation of the interviewer variance, with the highest recorded relative percentage bias varying from -5% for the 8% change scenario to -62% for the 92% change scenario. However, if the correct profile is chosen, w_{ij}= (0.5, 0.5) for W_{ij}=(0.5, 0.5), there does not seem to be any difference in the percentage relative bias across the scenarios with different interviewer change

proportions. In these scenarios accurate estimates are obtained both when choosing the correct weights or adjacent weights. It is only for the scenario with 92% interviewer changes that choosing the correct weights w_{ij}= (0.5, 0.5) results in a substantial improvement in the accuracy of the estimator compared to the adjacent weights w_{ij}=(0.6, 0.4) or w_{ij}=(0.4, 0.6). This trend of higher biases for scenarios with higher percentage of cases experiencing interviewer change is also observed for other Type A (Table 4) and Type B change profile scenarios (Table 5).

Table IV.3: Relative Percentage Bias for Type A, N=5760, N_p^I=240, σ_u^2=0.3, π=0.8, W_{ij}= (0.5, 0.5) Scenarios with Varying Percentage Change

	Interviewer Change				
w_{ij}	8%	21%	33%	50%	92%
0.9, 0.1	-4.84	-13.27	-21.72	-32.91	-60.92
0.8, 0.2	-2.40	-7.41	-12.86	-20.61	-43.81
0.7, 0.3	-0.58	-2.87	-5.61	-9.68	-24.37
0.6, 0.4	0.54	0.06	-0.84	-2.01	-7.89
0.5, 0.5	0.91	1.05	0.76	0.76	-1.37
0.4, 0.6	0.52	0.05	-1.02	-1.96	-8.25
0.3, 0.7	-0.60	-2.84	-5.92	-9.63	-25.00
0.2, 0.8	-2.41	-7.39	-13.28	-20.55	-44.52
0.1, 0.9	-4.87	-13.22	-22.23	-32.84	-61.55
DIC based	1.01	1.11	0.67	0.71	-1.74

Generally, for the W_{ij}=(0.5, 0.5) scenarios symmetry in the distribution of the absolute values of the biases across the models with different weight specifications can be observed, with the lowest bias obtained for the model specifying the correct model weights [w_{ij}=(0.5, 0.5)], and the highest biases obtained for the most unequally distributed and incorrect model weights [w_{ij}= (0.9, 0.1) or w_{ij}= (0.1, 0.9)]. These results are expected since models with w_{ij}= (0.9, 0.1) and models with w_{ij} = (0.1, 0.9) have the same degree of misspecification. However, for Type B, Type C and Type E scenarios some skewness in the distribution of the biases can be observed. Interestingly, one common feature for these interviewer change type profiles is a discrepancy in

the number of interviewers and their caseload between the previous and current waves. In contrast, Type A, Type D and Type F include 240 interviewers with a caseload of 24 households at both waves wave.

Scenarios with 50% change and w_{ij}=(0.9, 0.1) tend to show a mean relative percentage bias which is positive and close to zero for the model with the correct weights (w_{ij}=0.9, 0.1), continues to be positive and increases for more equal w_{ij}, then reduces, turns negative and then increases in magnitude as the model weights get closer to w_{ij}=(0.1, 0.9) (Table 6). The point at which the bias turns negative varies by change profile type. Though low biases are observed where the positive bias turns negative, sometimes matching the bias obtained for the model with the correct weights, the average DIC (presented in Appendix I) consistently shows higher values with greater discrepancies between the w_{ij} and the W_{ij}.

Table IV.4: Relative Percentage Bias for Type A, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240						N=2880; N_p^I=120			
	8% change			50% change			8% change		50% change	
						W_{ij}				
w_{ij}	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	-4.8	-2.4	1.3	-32.9	-16.6	1.2	-2.6	3.8	-35.4	-0.1
0.8, 0.2	-2.4	-0.7	2.3	-20.6	-6.4	8.4	-0.1	4.9	-23.1	7.1
0.7, 0.3	-0.6	0.4	2.6	-9.7	0.9	11.1	1.8	5.2	-12.1	9.7
0.6, 0.4	0.5	0.7	2.2	-2.0	3.6	8.0	2.8	4.7	-4.5	6.5
0.5, 0.5	0.9	0.3	1.0	0.8	0.6	-1.1	3.2	3.5	-1.9	-2.9
0.4, 0.6	0.5	-0.9	-1.0	-2.0	-7.8	-15.0	2.8	1.5	-4.7	-17.1
0.3, 0.7	-0.6	-2.7	-3.6	-9.6	-19.9	-31.3	1.6	-1.2	-12.4	-33.7
0.2, 0.8	-2.4	-5.2	-6.8	-20.6	-33.9	-47.7	-0.4	-4.5	-23.4	-50.3
0.1, 0.9	-4.9	-8.3	-10.5	-32.8	-47.7	-62.5	-2.9	-8.4	-35.8	-64.7
DIC based	1.0	0.7	2.5	0.7	1.2	4.6	3.8	5.7	-1.6	4.9

Table IV.5: Relative Percentage for Type B, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N= 5760; N_p^I=240				N=2880; N_p^I=120			
	8% change		50% change		8% change		50% change	
				W_{ij}				
w_{ij}	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	-5.0	0.3	-37.2	1.3	-7.2	0.8	-39.2	-0.9
0.8, 0.2	-3.0	1.7	-27.7	11.5	-5.2	2.2	-29.9	7.6
0.7, 0.3	-1.2	2.8	-17.1	20.6	-3.3	3.3	-19.6	15.0
0.6, 0.4	0.3	3.6	-7.1	25.9	-1.9	4.1	-9.9	19.3
0.5, 0.5	1.5	3.9	-0.3	25.0	-0.7	4.5	-3.5	18.6
0.4, 0.6	2.2	3.8	1.0	17.2	-0.0	4.4	-2.7	12.1
0.3, 0.7	2.5	3.2	-3.9	3.3	0.2	3.8	-8.0	0.6
0.2, 0.8	2.2	2.1	-13.6	-14.5	-0.1	2.7	-18.1	-12.4
0.1, 0.9	1.4	0.5	-26.4	-34.1	-0.9	1.0	-30.8	-28.6
DIC based	1.9	2.2	-1.0	8.4	0.2	3.9	-4.2	8.6

A worrying trend noticeable for the Type B change profile, 8% change, W_{ij}=(0.9, 0.1) scenarios (Table 4) and the Type F, 50% change, W_{ij}=(0.9, 0.1) scenarios (Table 6) is that there is some kind of symmetry in the biases for different weight profiles around w_{ij}=(0.5, 0.5) [usually noticeable only for W_{ij}=(0.5, 0.5) since the degree of misspecification is symmetrical around w_{ij}=(0.5, 0.5)]. In this case, the symmetry would be indicating that the model is identifying w_{ij}=(0.9, 0.1) and w_{ij}=(0.1, 0.9) models as fitting the data equally well. These results suggest that for unequally distributed real weights data, in cases of low percentage of multiple memberships or a very restrictive change profile, there may be insufficient information to correctly apportion the variance across two waves. The variance for cases with interviewer changes is $(w_{ij_p}{}^2 + w_{ij_c}{}^2)\sigma_u^2$ and both $(0.9^2 + 0.1^2)\sigma_u^2$ and $(0.1^2 + 0.9^2)\sigma_u^2$ will give the same estimated variance. For the Type B change profile, 8% change scenario the mean DIC increases slightly as the weight specification is changed from (0.9, 0.1) to (0.1, 0.9). On the other hand for the Type F, 50% change scenario the DIC across the various models is relatively stable, but showing a slightly higher DIC value for the (0.5, 0.5) model compared to models with more unequal weights. The Brooks-Draper and Raftery-Lewis diagnostics for these models

do not indicate any convergence problems. A similar pattern, of models with unequal weights (irrespective of whether the previous or current interviewers are given the greatest weight) performing better than models with equal weights, is observed in the first paper.

The comparison in bias between N=5760 and N=2880 scenarios shows a lot of small fluctuations (Table 4 & Table 5), generally showing non-substantial differences in the bias across the two sample size scenarios. The relatively large sample size for the N=2880 scenarios may explain this lack of effect of N on the estimator bias. Therefore, the result obtained does not exclude an effect of N on bias for smaller sample sizes.

A lot of variation in bias across different change profile types can be observed for the models including the most incorrect w_{ij} (Table 6). The reasons underlying this variation are not quite clear. The most important observation is that irrespective of the unequal number of interviewers and caseload across the two waves and the level of restriction in the allocation of change cases (random allocation or swapping between interviewers) the bias of the estimator across the different change profile types is relatively stable for the models including the correct and neighbouring w_{ij}.

A low relative percentage bias is obtained when the w_{ij} choice is based on the DIC. For any scenario basing the weights selection on the DIC is the best strategy, as it avoids the possibility of huge biases in the interviewer variance if weights are gravely misspecificed. Moreover, for equally distributed real MM weights [W_{ij}=(0.5, 0.5)], for all change profile types except Type F, getting the weights right on a theoretical basis [w_{ij}=(0.5, 0.5)] does not offer a major improvement in terms of the estimator bias compared to choosing the weights on the basis of the DIC. In contrast, substantially higher biases are obtained for the models including weights based on the DIC compared to the models including the correct theoretical weights profile (w_{ij}=0.9, 0.1) for W_{ij}=(0.9, 0.1). In fact, a clear trend for DIC-based weights is the higher bias for more unequally distributed W_{ij} for scenarios with a high percentage of cases with multiple interviewer memberships. Therefore, this trend is not observed for the 8% change scenarios. However, the absolute value for the

random effect estimator bias never exceeds 10% for the DIC-based weights models, in contrast with biases that exceed 60% for models with theoretically-based misspecified weights for the scenarios considered.

Table IV.6: Relative Percentage Bias for N=5760, N_p^I=240, 50% change, σ_u^2=0.3, π=0.8 with varying W_{ij} and Change Type Profile

			W_{ij}=0.5 0.5			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	-32.9	-37.2	-27.1	-30.8	-17.1	-18.6
0.8, 0.2	-20.6	-27.7	-15.8	-21.6	-9.4	-11.7
0.7, 0.3	-9.7	-17.1	-6.5	-11.5	-3.0	-6.1
0.6, 0.4	-2.0	-7.1	-0.6	-2.9	1.0	-2.4
0.5, 0.5	0.8	-0.3	0.5	0.5	1.1	-1.1
0.4, 0.6	-2.0	1.0	-4.4	-2.9	-3.5	-2.4
0.3, 0.7	-9.6	-3.9	-14.2	-11.5	-11.7	-6.1
0.2, 0.8	-20.6	-13.6	-23.3	-21.6	-20.9	-11.7
0.1, 0.9	-32.8	-26.4	-36.1	-30.8	-29.8	-18.5
DIC based	0.7	-1.0	-0.2	-0.9	-1.0	-6.1
			W_{ij}=0.9 0.1			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	1.2	1.3	0.6	0.1	1.3	0.5
0.8, 0.2	8.4	11.5	8.6	10.4	10.0	9.1
0.7, 0.3	11.1	20.6	13.4	19.3	16.4	16.2
0.6, 0.4	8.0	25.9	12.8	23.3	17.9	20.9
0.5, 0.5	-1.1	25.0	3.5	18.3	10.6	22.5
0.4, 0.6	-15.0	17.2	-16.4	3.3	-5.7	20.9
0.3, 0.7	-31.3	3.3	-39.2	-16.4	-23.4	16.2
0.2, 0.8	-47.7	-14.5	-55.6	-33.3	-36.4	9.1
0.1, 0.9	-62.5	-34.1	-65.9	-45.0	-45.2	0.6
DIC based	4.6	8.4	3.8	6.3	4.8	7.1

Here the parameter estimate chosen is the posterior mean. Celeux et al. (2006) suggest also considering the posterior mode or median. In this study the posterior median has been saved and its corresponding bias has been calculated. The biases for the posterior mean and the posterior median are almost identical and show the same trends with changes in the factors. These results are included in Appendix J.

To summarise, the main points on the relative percentage bias are the following:

- Negligible or low bias is observed for models specifying the correct model weights. Consequently, no trend across simulation factor values in the estimator bias for the models including correct model weights can be observed.
- Estimators of models with incorrect model weights show non-negligible, at times extremely high, bias.
- For models with incorrect weights higher biases are observed for scenarios with a greater percentage of cases experiencing interviewer change (proportion of cases with multiple memberships).
- As expected scenarios with (0.5, 0.5) real weights data show symmetry in the absolute biases around the (0.5, 0.5) weights model. Some skewness is observed for change profile types with unequal numbers of interviewers and unequal workloads across the two waves.
- For (0.9, 0.1) real weights scenarios including a larger number of cases with multiple memberships (50% change) the bias is positive, increases in effect size, then decreases and turns negative with greater misspecification in the model weights.
- For some scenarios with real multiple membership weights of (0.9, 0.1) symmetry in the biases is observed across the different models with different weights. For these scenarios there seems to be insufficient information for the total variance to be correctly apportioned across the two waves.
- No effect of halving the total sample size on the bias is noticeable for the sample sizes considered (N=5760 and N=2880).
- For (0.5, 0.5) real weights data basing the model weights on the DIC results in equally accurate estimates in comparison to models including correct theoretically-based weights.
- Across all possible scenarios acceptable levels of bias (less than 10%) are obtained when the model weight choice is based on the DIC compared to models with incorrect theoretical weights (up to 60%).

IV.5.2. Power

The null hypothesis specifies the true interviewer variance value to be zero. The power is equal to 1 in most scenarios across all w_{ij} specifications, and therefore less influenced by factor changes in comparison to other properties. Some exceptions are observed for very badly misspecified w_{ij}, especially for scenarios with high percentage changes and small N. For scenarios with 50% change and N=5760 the power is always above 0.90 across different change profile types, percentage change, W_{ij} and w_{ij}. The power of the Wald test across scenarios varying in terms of the percentage change for Type A, W_{ij}=(0.5, 0.5) scenarios (Table 7) is stable. It is only for the worst misspecified w_{ij} models of the 92% change scenario that substantially lower power is observed, indicating that only very high percentage change values have an influence on power. For the models with the most erroneous weights, that is w_{ij}=(0.9, 0.1) or w_{ij}=(0.1, 0.9), the power goes down to 0.87 for 92% change scenario.

Table IV.7: Power for Type A, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

w_{ij}	8% change	21% change	33% change	50% change	92% change
0.9, 0.1	1	1	1	1	0.87
0.8, 0.2	1	1	1	1	0.96
0.7, 0.3	1	1	1	1	0.98
0.6, 0.4	1	1	1	1	0.99
0.5, 0.5	1	1	1	1	0.99
0.4, 0.6	1	1	1	1	0.99
0.3, 0.7	1	1	1	1	0.98
0.2, 0.8	1	1	1	1	0.94
0.1, 0.9	1	1	1	1	0.87
DIC based	1	1	1	1	1

As expected, the scenarios with N=2880 (considered for Type A and Type B change profiles) show some lower values for the power of the Wald test in comparison to equivalent scenarios with N=5760. Table 8 shows that for

Type A scenarios a value below 0.90 for the power of the Wald tests is obtained for scenarios with N=2880, 50% change and W_{ij}=(0.9, 0.1) for the models which have the greatest degree of weights misspecification [w_{ij}=(0.3, 0.7), w_{ij}=(0.2, 0.8) and w_{ij}=(0.1, 0.9) models]. For Type B scenarios low power is obtained for models with misspecified weights for scenarios with N=2880 and 50% change for both W_{ij}=(0.5, 0.5) and W_{ij}=(0.9, 0.1) scenarios (Table 9). Therefore, the effect of N is only noticeable for high percentage change values and noticeable for different W_{ij} for different change profile type scenarios. Higher power is obtained for W_{ij} =(0.5, 0.5) scenarios in comparison to W_{ij}=(0.9, 0.1) scenarios for the Type A change profile for N=2880, 50% change scenarios. The opposite is true for the Type B change profile for N=2880, 50% change scenarios.

Table IV.8: Power for Type A, σ_u^2 =0.3, π =0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240						N=2880; N_p^I=120			
	8% change			50% change			8% change		50% change	
						W_{ij}				
w_{ij}	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	1	1	1	1	1	1	1	1	0.93	1.00
0.8, 0.2	1	1	1	1	1	1	1	1	0.96	1.00
0.7, 0.3	1	1	1	1	1	1	1	1	0.97	0.99
0.6, 0.4	1	1	1	1	1	1	1	1	0.98	0.99
0.5, 0.5	1	1	1	1	1	1	1	1.00	0.98	0.98
0.4, 0.6	1	1	1	1	1	1	1	1	0.98	0.94
0.3, 0.7	1	1	1	1.00	1.00	1.0	1	1	0.98	0.89
0.2, 0.8	1	1	1	1.00	1.00	0.98	1	1.00	0.96	0.75
0.1, 0.9	1	1	1	1.00	1.00	0.91	1	1.00	0.94	0.57
DIC based	1	1	1	1	1	1	1.00	1	0.99	1

A value of 1.00 represents a rounded up value of 1, whereas when a value of 1 indicates that all 1000 scenarios the null hypothesis is rejected

Table IV.9: Power for Type B, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240				N=2880; N_p^I=120			
	8% change		50% change		8% Change		50% Change	
					W_{ij}			
w_{ij}	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	1	1	1.00	1	1.00	1.00	0.89	1.00
0.8, 0.2	1	1	1.00	1	1.00	1.00	0.90	1.00
0.7, 0.3	1	1	1.00	1	1.00	1.00	0.92	1.00
0.6, 0.4	1	1	1.00	1	1.00	1.00	0.92	0.99
0.5, 0.5	1	1	1.00	1	1.00	1.00	0.93	0.99
0.4, 0.6	1	1	1.00	1	1.00	1.00	0.93	0.99
0.3, 0.7	1	1	1.00	1	1.00	1.00	0.91	0.97
0.2, 0.8	1	1	1.00	1.00	1.00	1.00	0.88	0.93
0.1, 0.9	1	1	0.99	0.99	1.00	1.00	0.83	0.83
DIC based	1	1	1	1	1.00	1.00	0.96	1.00

For the DIC-based weights models the power is greater than 0.95, and therefore in terms of power basing the weights on the DIC always yields good results. Since power values for DIC-based weights models are all optimal no pattern for power across different factors can be identified.

Table IV.10: Power for N=5760, N_p^I=240, 50% change, σ_u^2=0.3, π=0.8 with varying W_{ij} and Change Type Profile

			W_{ij}=(0.5, 0.5)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	1	1.00	1.00	1.00	1	1
0.8, 0.2	1	1.00	1	1.00	1	1
0.7, 0.3	1	1.00	1	1.00	1	1
0.6, 0.4	1	1.00	1	1.00	1	1
0.5, 0.5	1	1.00	1	1.00	1	1
0.4, 0.6	1	1.00	1	1	1	1
0.3, 0.7	1.00	1.00	1	1	1	1
0.2, 0.8	1.00	1.00	1	1	1	1
0.1, 0.9	1.00	0.99	1	1.00	1	1
DIC based	1	1.00	1	1	1	1
			W_{ij}=(0.9, 0.1)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	1	1	1	1	1	1
0.8, 0.2	1	1	1	1	1	1
0.7, 0.3	1	1	1	1	1	1
0.6, 0.4	1	1	1	1	1	1
0.5, 0.5	1	1	1	1	1	1
0.4, 0.6	1	1	1.00	1	1	1
0.3, 0.7	1.00	1.00	0.99	1	1	1
0.2, 0.8	0.98	1.00	0.98	1.00	1.00	1
0.1, 0.9	0.91	0.98	0.96	0.93	1	1
DIC based	1	1	1	1	1	1

To summarise, the main points on the power of the Wald test are the following:

- The power is equal to 1 in most scenarios across all model weight specifications.
- Power is less influenced by factor value changes in comparison to the properties of the variance estimator across all model weight specifications.
- Lower values of power are obtained for very badly misspecified weights models for scenarios with a high proportion of cases being associated with multiple memberships and small total sample sizes.

- The models including weights based on the DIC always obtain optimal power values (greater than 0.95).

IV.5.3. Confidence Interval Coverage

While most models with correct w_{ij} obtain a confidence interval coverage rate close to the nominal 95% rate, confirming the result presented by Browne et al. (2001) for their simulated example, some correctly specified models obtain slightly lower rates. However, these rates do not fall below 90% for the scenarios considered. In contrast to Browne et al. (2001) which base the coverage rate on the 95% credible confidence interval from the MCMC chains, the results presented in this study are based on the 95% Wald confidence interval. The 95% credible intervals have also been assessed, but will not be presented in this paper. Both results are included in Appendix K. The two measures show similar values and the same trends across factors. However, for models with the worst w_{ij} for some scenarios obtaining low coverage rates (especially for scenarios with a high proportion of multiple memberships and small sample size) the 95% credible interval performs just slightly better than the 95% Wald confidence interval.

Coverage rates which differ substantially from the nominal value of 95% are observed for misspecified weights models. With increasing multiple memberships (percentage change) the confidence interval coverage rate goes down for misspecified models (Table 11–Table 13). With increasing values of percentage change for scenarios with a Type A profile, W_{ij}=(0.5, 0.5), and typical values on other factors the confidence interval coverage for models with the most incorrect weights [w_{ij}=(0.1, 0.9) or w_{ij}=(0.9, 0.1)] decreases, from around 92% for the 8% change scenario to 4% for the 92% scenario. Again when comparing other Type A and Type B scenarios varying in the percentage change in Tables 12 and 13, keeping constant other factors, it can be noticed that generally the confidence interval coverage of estimators for scenarios with a 50% change is lower than that for the 8% change scenarios.

Table IV.11: CI coverage for Type A, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change

w_{ij}	8% change	21% change	33% change	50% change	92% change
0.9, 0.1	92.1	81.6	71.0	44.1	4.4
0.8, 0.2	93.8	88.1	85.7	74.5	33.1
0.7, 0.3	94.1	92.0	93.0	88.6	73.1
0.6, 0.4	94.3	93.2	94.7	93.5	91.1
0.5, 0.5	94.6	94.2	95.2	94.7	92.7
0.4, 0.6	94.4	93.7	94.4	94.4	90.0
0.3, 0.7	93.7	91.8	91.9	89.9	72.6
0.2, 0.8	93.4	88.3	84.7	73.5	30.6
0.1, 0.9	92.1	83.5	70.7	45.1	3.9
DIC based	94.6	94.0	94.9	94.4	92.1

The effect of sample size (N) on the confidence interval coverage is not so intuitive. There is some indication that for models with theoretical weights for scenarios with 50% change the confidence interval coverage is higher for N=2880 scenarios compared to N=5760 scenarios for misspecified models and slightly lower for models with correct w_{ij} (Tables 12 & 13). So for example for the Type B, 50% change, W_{ij}=(0.5, 0.5) case coverage rates of 57.6% and 90.9% for the w_{ij}=(0.9, 0.1) and w_{ij}=(0.5, 0.5) models are obtained for the N=2880 scenarios, compared to 34.5% and 94.4% for the N=5760 scenarios (Table 13). For the 8% change scenarios no trend can be identified, indicating that the effect of N is only noticeable for data with a high percentage of multiple memberships. In fact for both change Type A and Type B scenarios confidence interval coverage rates do not fall below 87.5% for all possible weights specifications for 8% change.

The symmetry in the properties expected for W_{ij}=(0.5, 0.5) scenarios across the 9 theoretically-based weights models around the w_{ij}=(0.5, 0.5) model is not observed perfectly for Type B, Type C and Type E scenarios. This skewness has also been observed for the bias of the estimator for these same change profile scenarios. Interestingly, the average DIC value across the different models with different theoretically-based w_{ij} shows perfect symmetry

for all change profiles types, with the lowest DIC obtained for the w_{ij}=(0.5, 0.5) model and the highest values (approximately equal) for the w_{ij}=(0.1, 0.9) and the w_{ij}=(0.9, 0.1) models.

Table IV.12: CI coverage for Type A, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240						N=2880; N_p^I=120			
	8% change			50% change			8% change		50% change	
						W_{ij}				
w_{ij}	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	92.1	93.7	94.5	44.1	79.1	95.2	92.0	96.4	61.3	93.7
0.8, 0.2	93.8	94.4	94.6	74.5	90.4	95.4	92.4	96.4	79.5	94.9
0.7, 0.3	94.1	95.0	94.6	88.6	94.1	95.5	93.2	96.5	89.1	95.0
0.6, 0.4	94.3	94.8	94.4	93.5	94.4	95.6	93.6	96.6	92.3	94.6
0.5, 0.5	94.6	94.8	93.7	94.7	94.3	94.6	93.6	96.4	93.9	93.0
0.4, 0.6	94.4	94.4	93.6	94.4	90.8	83.4	93.8	95.7	93.7	85.7
0.3, 0.7	93.7	93.4	92.4	89.9	76.2	54.4	93.3	94.9	89.0	67.7
0.2, 0.8	93.4	92.3	90.4	73.5	45.4	15.7	92.9	92.9	79.6	39.2
0.1, 0.9	92.1	89.4	87.5	45.1	12.9	1.7	92.1	91.1	60.8	12.6
DIC based	94.6	94.8	94.6	94.4	93.9	96.0	93.8	96.5	93.3	94.8

In the case of W_{ij} =(0.9, 0.1) scenarios the coverage rates remain relatively stable or only decrease slightly when specifying the next couple of weight schemes in comparison to the correct weights. However, for the most erroneously specified weights [w_{ij}=(0.1, 0.9), w_{ij}=(0.2, 0.8) and w_{ij}=(0.3, 0.7)] much lower coverage rates are observed.

The coverage rates observed for the models with the most incorrect weights vary across different change profile types and between each W_{ij} factor value. However, the models specifying the correct weights and models specifying adjacent weights show similar confidence interval coverage rates across different profile types and W_{ij}. For the unequally distributed W_{ij} scenarios the change profile types including a higher number of total interviewers (480 interviewers for Type B, and 360 interviewers for Type D and

Type F) obtain better coverage rates than the change profile types with a total of 240 distinct interviewers (Type A, Type C and Type E) for the models with incorrect weights.

Table IV.13: CI coverage for Type B, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240				N=2880; N_p^I=120			
	8% change		50% change		8% change		50% change	
				W_{ij}				
w_{ij}	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	90.6	94.3	34.5	95.9	90.6	94.9	57.6	92.8
0.8, 0.2	92.5	94.7	60.2	96.1	92.3	95.6	71.8	94.3
0.7, 0.3	93.6	95.1	80.4	91.7	93.2	95.4	82.7	94.3
0.6, 0.4	94.1	95.5	91.9	87.0	93.6	95.5	88.8	93.6
0.5, 0.5	94.5	95.3	94.4	87.9	94.0	95.4	90.9	93.6
0.4, 0.6	94.7	95.5	94.3	92.9	94.1	95.7	91.6	93.8
0.3, 0.7	94.6	95.1	93.3	95.0	93.9	95.4	90.6	93.7
0.2, 0.8	94.8	94.8	86.2	85.2	94.0	95.0	84.9	86.5
0.1, 0.9	94.5	94.6	66.0	50.6	93.7	95.0	72.5	73.0
DIC based	94.4	95.0	92.6	94.8	94.2	95.3	89.2	94.7

The models specifying the correct weights do not offer a substantial improvement on the confidence interval coverage of the estimator over models with weights based on the DIC. The only exception to this trend is the scenario with change profile Type F with W_{ij}=(0.5, 0.5), where the model including weights based on the DIC has a confidence interval coverage 87.8% compared to 92.7% for the model with w_{ij}=(0.5, 0.5) (Table 14). However, for this scenario coverage rates fall to 73% for models including incorrect theoretically-based weights. Therefore, in the case of the confidence interval coverage, relying on the DIC to select the model weights is the best strategy to avoid low coverage rates when the real multiple membership weights are unknown.

Table IV.14: CI coverage for N=5760, N_p^I=240, 50% change, σ_u^2=0.3, π=0.8 with varying W_{ij} and Change Type Profile

			W_{ij}=(0.5, 0.5)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	44.1	34.5	57.7	49.3	76.9	73.3
0.8, 0.2	74.5	60.2	79.0	72.4	88.2	85.1
0.7, 0.3	88.6	80.4	89.8	87.6	93.0	90.0
0.6, 0.4	93.5	91.9	93.0	93.1	94.8	92.0
0.5, 0.5	94.7	94.4	93.2	93.6	94.7	92.7
0.4, 0.6	94.4	94.3	90.9	92.1	93.2	92.2
0.3, 0.7	89.9	93.3	81.7	86.5	85.1	90.1
0.2, 0.8	73.5	86.2	66.1	72.1	73.9	84.9
0.1, 0.9	45.1	66.0	37.3	50.2	54.0	73.3
DIC based	94.4	92.6	92.6	92.2	92.3	87.8
			W_{ij}=(0.9, 0.1)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	95.2	95.9	94.6	93.8	94.4	94.2
0.8, 0.2	95.4	96.1	94.9	94.1	94.4	94.8
0.7, 0.3	95.5	91.7	94.4	90.8	92.2	93.0
0.6, 0.4	95.6	87.0	94.4	89.4	92.5	90.8
0.5, 0.5	94.6	87.9	93.6	91.7	93.0	90.0
0.4, 0.6	83.4	92.9	82.8	94.6	89.2	90.8
0.3, 0.7	54.4	95.0	41.7	81.1	69.3	93.2
0.2, 0.8	15.7	85.2	6.1	47.5	42.2	94.9
0.1, 0.9	1.7	50.6	0.3	19.1	16.9	94.3
DIC based	96.0	94.8	94.7	93.3	93.9	94.3

To summarise, the main points on the confidence interval coverage are the following:

- Models specifying the correct model weights obtain confidence interval coverage rates close to the nominal 95% rate, with the lowest rate obtained for the scenarios considered being higher than 90%. Consequently, the confidence interval coverage rates for the models with the correct or neighbouring weights do not vary by simulation factor.
- Extremely low coverage rates (even below 5%) are obtained for models with very badly misspecified weights.

- The lowest confidence interval coverage rates are observed for scenarios with a high percentage of cases with multiple memberships.
- For models with weights based on theory halving the sample size results in slightly lower coverage rates for correctly specified models and higher coverage rates for incorrect weights models. This effect of decreasing the sample size is only noticeable for data with a high percentage (50%) of cases with multiple memberships.
- As expected, scenarios with (0.5, 0.5) real weights data show symmetry in the confidence interval coverage around the (0.5, 0.5) weights model. Some skewness is observed for change profile types with unequal numbers of interviewers and unequal workloads across the two waves.
- Scenarios with (0.9, 0.1) real weights obtain relatively high and stable coverage rates when specifying the next couple of weight schemes in comparison to the correct weights. However, much lower coverage rates are observed for the models including the most erroneously specified weights.
- For scenarios with (0.9, 0.1) real weights the change profile types including a higher number of total interviewers obtain better coverage rates for the models with incorrect weights.
- The models with weights based on the DIC obtain equally high coverage rates as models including the correct model weights for most scenarios.

IV.5.4. Standard Error

When comparing the standard error properties of the variance estimator for models with different w_{ij} a clear pattern is identified. Higher standard errors for the models specifying equal w_{ij}, irrespective of whether these are the correct weights, are observed. A trend with an increase in the proportion of data with multiple memberships can be observed for the standard errors. Table 15 shows that for scenarios with typical factor values and W_{ij}=(0.5, 0.5), Type A profiles for models with equal w_{ij}, the standard error decreases for incorrect w_{ij} ($w_{ij} \neq W_{ij}$) and increases for correct w_{ij} ($w_{ij} = W_{ij}$) with increasing proportions of interviewer changes. This trend can also be observed for other

Type A and Type B scenarios with W_{ij} =(0.5, 0.5) (Tables 16 & 17). Consequently, the underestimation of incorrectly specified w_{ij} models is greater for scenarios with a higher percentage of multiple membership cases.

Table IV.15: Standard Error for Type A, N=5760, N_p^I=240, σ_u^2=0.3, π=0.8, W_{ij}= (0.5, 0.5) Scenarios with Varying Percentage Change

w_{ij}	8% change	21% change	33% change	50% change	92% change
0.9, 0.1	0.054	0.052	0.050	0.048	0.042
0.8, 0.2	0.056	0.055	0.054	0.054	0.053
0.7, 0.3	0.056	0.057	0.058	0.059	0.064
0.6, 0.4	0.057	0.058	0.060	0.063	0.073
0.5, 0.5	0.057	0.059	0.061	0.064	0.076
0.4, 0.6	0.057	0.058	0.060	0.063	0.073
0.3, 0.7	0.056	0.057	0.058	0.059	0.064
0.2, 0.8	0.056	0.055	0.054	0.054	0.053
0.1, 0.9	0.054	0.052	0.050	0.048	0.042
DIC based	0.057	0.059	0.060	0.063	0.075

Table IV.16: Standard Error for Type A, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240						N=2880; N_p^I=120			
	8% change			50% change			8% change		50% change	
	W_{ij}									
w_{ij}	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.7, 0.3	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	0.054	0.055	0.056	0.048	0.053	0.058	0.080	0.082	0.068	0.084
0.8, 0.2	0.056	0.056	0.057	0.054	0.058	0.063	0.081	0.083	0.077	0.090
0.7, 0.3	0.056	0.057	0.057	0.059	0.062	0.065	0.082	0.084	0.085	0.094
0.6, 0.4	0.057	0.057	0.057	0.063	0.064	0.066	0.083	0.084	0.090	0.094
0.5, 0.5	0.057	0.057	0.057	0.064	0.064	0.063	0.083	0.083	0.091	0.091
0.4, 0.6	0.057	0.056	0.056	0.063	0.061	0.058	0.083	0.082	0.090	0.084
0.3, 0.7	0.056	0.056	0.055	0.059	0.056	0.052	0.082	0.081	0.084	0.075
0.2, 0.8	0.056	0.055	0.054	0.054	0.050	0.045	0.081	0.079	0.077	0.064
0.1, 0.9	0.054	0.053	0.052	0.048	0.043	0.039	0.079	0.077	0.068	0.053
DIC based	0.057	0.057	0.057	0.063	0.062	0.060	0.083	0.084	0.090	0.087

For scenarios with a low percentage of cases with multiple memberships the standard error values across the different models with theoretical weights are relatively stable. For scenarios with a higher percentage of cases with multiple memberships the standard error values vary across the different models with different weights.

Table IV.17: Standard Error for Type B, σ_u^2=0.3, π=0.8 Scenarios with Varying Percentage Change, N, N_p^I and W_{ij}

	N=5760; N_p^I=240				N=2880; N_p^I=120			
	8% change		50% change		8% change		50% change	
				W_{ij}				
w_{ij}	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1	0.5, 0.5	0.9, 0.1
0.9, 0.1	0.054	0.056	0.047	0.059	0.077	0.081	0.067	0.083
0.8, 0.2	0.055	0.057	0.053	0.065	0.079	0.082	0.076	0.090
0.7, 0.3	0.056	0.058	0.059	0.070	0.080	0.083	0.085	0.097
0.6, 0.4	0.057	0.058	0.065	0.073	0.082	0.084	0.093	0.101
0.5, 0.5	0.058	0.059	0.068	0.074	0.082	0.085	0.098	0.101
0.4, 0.6	0.058	0.059	0.068	0.071	0.083	0.085	0.098	0.098
0.3, 0.7	0.058	0.059	0.066	0.066	0.083	0.085	0.094	0.092
0.2, 0.8	0.058	0.058	0.061	0.060	0.083	0.084	0.087	0.086
0.1, 0.9	0.058	0.058	0.055	0.053	0.082	0.083	0.079	0.078
DIC based	0.058	0.057	0.067	0.063	0.082	0.083	0.094	0.090

For W_{ij}=(0.9, 0.1) scenarios the models with the correct weights have standard errors equal to or higher than the model with the worst specified weights, but lower standard errors than the models with equally distributed w_{ij}. For W_{ij}=(0.5, 0.5) scenarios the distribution of the standard errors across the different models with different w_{ij} is symmetrical, with the model specifying the correct weights having the highest standard errors. Models with misspecified w_{ij} therefore have smaller standard errors. Some skewness is observed for change profile Type B. As expected, halving N results in higher standard errors for all weight profiles (Tables 16 & 17). With increasing values

of percentage change and smaller N the standard error difference between the models with the worst misspecified weights and the correct weights increases.

Table IV.18: Standard Error for N=5760, N_p^I=240, 50% change, σ_u^2=0.3, π=0.8 with varying w_{ij} and Change Type Profile

			W_{ij}=(0.5, 0.5)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	0.048	0.047	0.049	0.048	0.052	0.052
0.8, 0.2	0.054	0.053	0.054	0.054	0.056	0.056
0.7, 0.3	0.059	0.059	0.059	0.060	0.061	0.060
0.6, 0.4	0.063	0.065	0.064	0.066	0.065	0.063
0.5, 0.5	0.064	0.068	0.066	0.068	0.067	0.064
0.4, 0.6	0.063	0.068	0.066	0.066	0.066	0.063
0.3, 0.7	0.059	0.066	0.061	0.060	0.061	0.060
0.2, 0.8	0.054	0.061	0.054	0.054	0.055	0.056
0.1, 0.9	0.048	0.055	0.046	0.048	0.049	0.052
DIC based	0.063	0.067	0.065	0.066	0.065	0.060
			W_{ij}=(0.9, 0.1)			
w_{ij}	Type A	Type B	Type C	Type D	Type E	Type F
0.9, 0.1	0.058	0.059	0.058	0.058	0.059	0.058
0.8, 0.2	0.063	0.065	0.064	0.065	0.065	0.064
0.7, 0.3	0.065	0.070	0.069	0.071	0.072	0.069
0.6, 0.4	0.066	0.073	0.074	0.076	0.078	0.072
0.5, 0.5	0.063	0.074	0.075	0.076	0.080	0.073
0.4, 0.6	0.058	0.071	0.069	0.071	0.072	0.072
0.3, 0.7	0.052	0.066	0.054	0.061	0.060	0.069
0.2, 0.8	0.045	0.060	0.042	0.050	0.049	0.064
0.1, 0.9	0.039	0.053	0.034	0.043	0.042	0.058
DIC based	0.060	0.063	0.060	0.062	0.061	0.062

Very little difference in the standard errors across different change profile types for the W_{ij}=(0.5, 0.5) scenarios can be observed. In comparison, the W_{ij}=(0.9, 0.1) scenarios show standard errors which are closer in value to the standard errors of the models with the correct weights for the change profiles that include new interviewers in the current wave (Type B, Type D and Type F) for the models with the worst specified w_{ij}. Scenarios that allocate the change cases to existing interviewers (Type A, Type C and Type E) obtain

smaller standard errors for the models with the worst specified w_{ij}. The improved standard error properties for scenarios including new interviewers at the current wave may be due to the larger number of higher-level units.

The standard error of the estimator for the models including weights based on the DIC is approximately equal to the standard error for the models specifying the correct weights. The standard errors of the models including DIC-based weights are approximately equal across the different change profile types and different w_{ij}, higher for a greater proportion of interviewer changes and higher for smaller N. These results would indicate that an empirical assessment for the choice of DIC leads to good estimates of the standard error.

To summarise, the main points on the standard errors of the variance estimator are the following:

- The highest standard errors across models with different weight specifications are observed for models including equally distributed weights (irrespective of whether these are the correct weights).
- For (0.5, 0.5) real weights data, with increasing percentages of cases with multiple memberships the standard error of models with equally distributed weights decreases when the weights are incorrect and increases when the weights are correct.
- Larger standard errors are observed for scenarios with smaller total sample sizes.
- As expected, scenarios with (0.5, 0.5) real weights data show symmetry in the standard error values around the (0.5, 0.5) weights model. The correct model obtains the highest standard error.
- For (0.9, 0.1) real weights data the correct model has equal or higher standard errors than the model with the most incorrect weights but lower standard errors than the models with equally distributed weights.
- Very little difference in the standard errors across different change profile types for the (0.5, 0.5) real weights scenarios can be observed.
- For the (0.9, 0.1) real weights scenarios the smaller standard errors for models with incorrect weights is more pronounced for change profile

types which maintain the same pool of interviewers across the two waves (smaller number of higher-level units).

- The standard error for the models including weights based on the DIC is approximately equal to the standard error for the models specifying the correct weights.
- Models with weights based on DIC show larger standard errors for smaller sample sizes and higher percentages of multiple membership cases.

IV.5.5. DIC Reliability Measure

This section explores the reliability of the DIC in choosing the model with the correct multiple membership weights w_{ij}. The graph shows the frequency distribution of the w_{ij} specified for the 1000 models (out of the 9000 models of each scenario) corresponding with the lowest DIC. The tables show the proportion of these 1000 models that have the correct w_{ij} and the proportion which have the correct or adjacent w_{ij} for different scenarios.

First we examine the effect of the proportion of cases with multiple memberships on the DIC reliability. In Figure 1 it can be noticed that for Type A, W_{ij}=(0.5, 0.5) scenarios with varying degrees of percentage change with typical values for the other factors the DIC performs better for scenarios with a greater proportion of cases experiencing change. This is contrary to the results obtained for the properties of the variance estimator and the test statistic, reviewed above, which showed that worse estimator properties and power of the Wald test is obtained for scenarios with a greater percentage of cases experiencing interviewer change. The greatest increase in the correct model weights being selected, of 12.5%, is noticed when increasing the proportion of cases experiencing interviewer change from 8% to 21%.

For the scenarios included in Figure 1 the highest proportion of times that the lowest DIC corresponds to the w_{ij}=(0.5, 0.5) model is obtained for the scenario including multiple memberships in 92% of cases. This proportion amounts to only 49.5%, which would suggest that the DIC does not offer a very

precise measure for choosing the correct w_{ij}. However, if selecting either the correct weights or the next most precise weights [in the case of W_{ij}=(0.5, 0.5) this includes w_{ij}=(0.5, 0.5) or w_{ij}=(0.4, 0.6) or w_{ij}=(0.6, 0.4)] is deemed acceptable, the results are more encouraging. Even for scenarios with 21% interviewer changes an adequate weighting scheme is selected 75% of the time. This proportion goes up to 81.7%, 90.3% and 93.4% for the 33%, 50% and 92% interviewer change scenarios respectively. This trend of higher DIC reliability for higher multiple membership proportions is also observed for other Type A and Type B scenarios (Table 20). However, for Type B, W_{ij}=(0.9, 0.1) scenarios no difference is observed in the DIC reliability across different percentage change scenarios.

Figure IV.1: Frequency Distribution of the Model Weights for the DIC-based Weights Models for Type A, N=5760, N_p^I=240, σ_u^2=0.3, π=0.8, W_{ij}=(0.5, 0.5) Scenarios with Varying Percentage Change

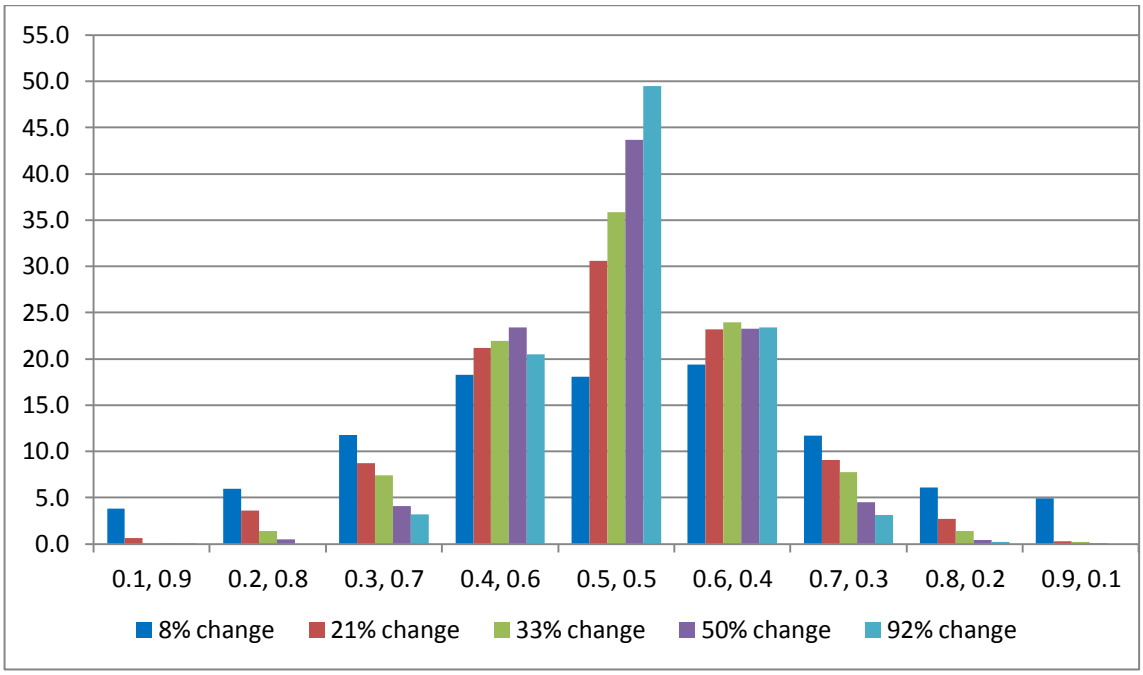

Table IV.19: DIC Reliability Measure for Type A, N=5760, N_p^I=240, σ_u^2=0.3, π=0.8, W_{ij}= (0.5, 0.5) Scenarios with Varying Percentage Change

Percentage Change	Proportion Correct Weights	Proportion Correct/ Adjacent Weights
8%	18.1	55.7
21%	30.6	75.0
33%	35.8	81.7
50%	43.7	90.3
92%	49.5	93.4

Table IV.20: DIC Reliability Measure for N_p^I=240, σ_u^2=0.3, π=0.8 Scenarios with Varying Profile Change Type, Percentage Change, N, and W_{ij}

Real Weights W_{ij}	Change Type	Sample Size	Percentage Change	Proportion Correct Weights	Proportion Correct/ Adjacent Weights
0.5, 0.5	A	5760	8%	18.1	55.7
			50%	43.7	90.3
		2880	8%	12.3	40.3
			50%	29.9	73.4
	B	5760	8%	13.2	39.1
			50%	32.7	80.5
		2880	8%	9.6	28.5
			50%	27.0	65.9
0.9, 0.1	A	5760	8%	57.9	76.0
			50%	70.2	93.4
		2880	8%	55.6	70.3
			50%	64.1	86.5
	B	5760	8%	58.0	71.7
			50%	60.8	83.8
		2880	8%	49.4	61.6
			50%	49.1	65.1
0.7, 0.3	A	5760	8%	16.2	49.8
			50%	49.8	89.4

For both Type A and Type B scenarios, halving the total sample size, and by consequence the number of interviewers, while maintaining the same

multiple membership proportions, results in drastic reductions in the ability of the DIC measure to correctly identify the correct w_{ij} for the W_{ij}=(0.5, 0.5) scenarios. However, the effect of N on the DIC reliability measure varies by change profile type and by W_{ij} (Table 20). In the case of the Type A scenarios with 50% interview changes the models specifying w_{ij}=(0.5, 0.5) correspond to the lowest DIC 13.8% fewer times for the N=2880 scenario compared to the N=5760 scenario. This reduction is less, more precisely 5.8%, for the 8% interviewer changes scenarios. For some reason not obvious to the author, these reductions are equal to the square root of the sample size ratios, that is $\sqrt{\frac{2880}{5760}}$. For the Type B scenarios this scaling effect is observed for the 8% change scenarios, but for the 50% change the factor is equal to $\frac{2880^{\frac{1}{4}}}{5760}$.

Halving N for W_{ij} =(0.9, 0.1) scenarios also impacts negatively the reliability of the DIC to correctly detect the MM weights. This DIC reliability decrease for smaller N is noticeable to a greater extent for the Type B scenarios for both percentage change profile types considered. This decrease happens to be equal to the scale of $\frac{2880^{\frac{1}{4}}}{5760}$. For the Type A scenarios some difference in the DIC reliability is observed for the 50% change scenarios but not the 8% change ones. The effect of N on the probability of the models with the correct or adjacent weights corresponding with the lowest DIC is more pronounced for Type B scenarios than Type A scenarios.

Therefore, a decrease in N has a greater impact when the W_{ij} are equal. Since the MM weights are usually unknown to the researcher, it is best to base the assessment of the DIC reliability for particular total sample sizes and interviewer change profiles on the equal W_{ij}. This approach will be a conservative one for cases with unequal W_{ij}. For W_{ij}=(0.9, 0.1) scenarios, Type B profiles seem to affected more than Type A scenarios by a change in N. For W_{ij}=(0.5, 0.5) scenarios Type B profiles are affected differently by a change in N for different percentage change scenarios, with higher percentage change scenarios showing a smaller decrease in DIC reliability with decreases in N. This shows how different scenario factors interact.

Next the impact of different W_{ij} for scenarios including different proportions of interviewer changes on the DIC reliability will be considered. If either the correct weight or adjacent weights are deemed accepted choices, there is not much discrepancy in the performance of the DIC in detecting the real MM weights for the scenarios including 50% change. Irrespective of the value of W_{ij} the DIC value is at its lowest value for the correct model weights or the next best weights around 90% of times for Type A scenarios and around 80% for Type B scenarios. The slightly higher proportions for W_{ij}=(0.9, 0.1) may simply be due to chance.

On the other hand, when the data includes 8% change the DIC reliability measure varies by W_{ij}. In Type A scenarios the correct multiple membership weights are detected more frequently (76% of times) for W_{ij}=(0.9, 0.1) than W_{ij}=(0.5, 0.5) (55.7%) and W_{ij}=0.7, 0.3 (49.8%) scenarios. Similarly, for Type B scenarios the correct multiple membership weights are detected more frequently (71.7% of times) for W_{ij} =(0.9, 0.1) than W_{ij} =(0.5, 0.5) (39.1%) scenarios. This higher DIC reliability for the W_{ij}=(0.9, 0.1) scenarios may be due to the boundary effect of this weighting scheme which only has one possible adjacent weighting scheme, w_{ij}=(0.8, 0.2), since w_{ij}=(1, 0) is not being considered as this simply represents a 2-level model.

If the strictest measure of DIC reliability is considered – equal to the proportion of times the correct weights correspond to the lowest DIC – a different result is obtained. In this case a change in the DIC reliability with different W_{ij} is observed for both the 8% and the 50% change scenarios. No substantial difference in the results for W_{ij} =(0.5, 0.5) and W_{ij} =(0.7, 0.3) scenarios are observed for Type A scenarios. A substantial increase in the DIC reliability is observed for the most unequal W_{ij} [W_{ij}=(0.9, 0.1)] for both Type A and Type B scenarios, confirming some kind of threshold effect in the influence of W_{ij} on the DIC measure performance. Moreover, the increase in the DIC reliability for more unequal W_{ij} is greater than for the previous DIC reliability definition. A bigger increase in the DIC reliability for more unequal W_{ij} can be observed for the 8% change scenarios compared to the 50% scenario.

This result confirms the earlier finding that for low MM proportions the influence of W_{ij} on DIC reliability is greater.

It is expected that the DIC performs better for situations where one interviewer is dominant compared to situations where the current and previous wave interviewers have equal influence since the former situation is closer to a purely hierarchical structure. For situations with one dominant interviewer the negative influence of a lack of interviewer change on the DIC's ability to identify the correct weights is less than for situations with two interviewers of equal influence.

Table IV.21: DIC Reliability Measure for N=5760, N_p^I=240, 50% change, σ_u^2=0.3, π=0.8 Scenarios with Varying Profile Change Type and W_{ij}

W_{ij}	Change Profile Type	Proportion Correct Weights	Proportion Correct/ Adjacent Weights
0.5, 0.5	A	43.7	90.3
	B	32.7	80.5
	C	37.8	84.9
	D	38.4	84.3
	E	29.6	72.2
	F	32.6	60.2
0.9, 0.1	A	70.2	93.4
	B	60.8	83.8
	C	71.2	94.5
	D	62.1	83.4
	E	68.8	91.2
	F	29.0	36.1

Table 21 compares the DIC reliability values for different change profile type scenarios with typical values for the other factors. The DIC reliability varies across comparable scenarios with different change profile types, performing the best for Type A scenarios and the worst for type F scenarios. In most cases, when comparing Type A to Type B, Type C to Type D and Type E to Type F the change profile types that do not include new interviewers in the current wave (Type A, Type C, Type E) fair better than the other change profile types. This result is more consistent for W_{ij}=(0.9, 0.1) scenarios. Therefore, to

the extent that new interviewers are introduced at the current wave to take on the workload for change cases, the DIC will be less useful as a method of detecting the correct weights. This result can be explained in terms of the greater amount of information available on each interviewer to identify interviewer effects when the same set of interviewers is maintained across both waves. In practice, it is more likely that the change cases are partly distributed to interviewers present in the previous wave and partly to newly recruited interviewers.

To summarise, the main points on the reliability of the DIC to identify the real weights are the following:

- DIC performs better for scenarios with a greater percentage of cases with multiple memberships.
- Halving the sample size results in drastic reductions in the reliability of the DIC. The change factor is dependent on the real weights and the profile change type.
- The reliability of the DIC varies by real weights, showing better results for unequally distributed weights data, noticeable to a greater extent for scenarios with a low percentage of cases with multiple memberships.
- The DIC performs better for change profile types that use the same pool of interviewers across both waves in comparison to change profile types that introduce new interviewers at the current wave.
- The DIC does not offer a very precise measure for choosing the exact correct model weights. However, the results are more encouraging when both the correct weights and the neighbouring weights are considered acceptable.

IV.5.6. Limited Pairing Scenarios

The model specification suggests that since there is simply one distribution for the interviewer random effects the pairings of multiple memberships, and how mixed the pairings are, is not a factor impacting the properties of the variance parameter and test statistic for the multiple membership model. This

deduction will be tested by comparing the results of scenarios where the change cases are allocated to another interviewer in the current wave randomly to two scenarios where change cases swap interviewer allocation between pairings and alternatively groups of four interviews in the current wave. An example scenario is presented in Table 22, listing the change indicator and the interviewer allocations for the previous and current waves for the three different interviewer swapping possibilities, to better help understand the different change profiles. This example scenario includes 16 cases allocated to 8 interviewers with 2 cases each in previous wave and 50% change.

Table IV.22: Interviewer Case Allocations for the Example Scenario

Case No	Change Indicator	P_Int Code	C_Int Code	P_Int Code	C_Int Code	P_Int Code	C_Int Code
		Random Allocation from Total Int Pool		Swapping between Pairs		Swapping between Groups of 4	
1	1	1	4	1	2	1	4
2	0	1	1	1	1	1	1
3	0	2	2	2	2	2	2
4	1	2	3	2	1	2	3
5	0	3	3	3	3	3	3
6	1	3	2	3	4	3	2
7	1	4	1	4	3	4	1
8	0	4	4	4	4	4	4
9	0	5	5	5	5	5	5
10	1	5	8	5	6	5	8
11	1	6	5	6	5	6	7
12	0	6	6	6	6	6	6
13	0	7	7	7	7	7	7
14	1	7	6	7	8	7	8
15	1	8	7	8	7	8	7
16	0	8	8	8	8	8	8

Table IV.23: Properties of the Estimator and Test Statistic and DIC Reliability Measure for Type A, 50% change, N=5760, N^I=240, W_{ij}=(0.5, 0.5), σ_u^2=0.3 and π=0.8 scenarios

w_{ij}	Relative Percentage Bias	CI Coverage	Power	Standard Error	DIC Reliability Measure
Random Allocation from Total Pool of Interviewers					
0.9, 0.1	-32.91	44.1	1	0.048	0.1
0.8, 0.2	-20.61	74.5	1	0.054	0.5
0.7, 0.3	-9.68	88.6	1	0.059	4.1
0.6, 0.4	-2.01	93.5	1	0.063	23.38
0.5, 0.5	0.76	94.7	1	0.064	43.66
0.4, 0.6	-1.96	94.4	1	0.063	23.28
0.3, 0.7	-9.63	89.9	0.999	0.059	4.5
0.2, 0.8	-20.55	73.5	0.999	0.054	0.4
0.1, 0.9	-32.84	45.1	0.998	0.048	0.1
DIC	0.71	94.4	1	0.063	
Swapping between Pairs					
0.9, 0.1	-29.74	51.8	1	0.048	
0.8, 0.2	-19.57	76.5	1	0.053	0.9
0.7, 0.3	-10.17	88.2	1	0.059	6.89
0.6, 0.4	-3.16	93.8	1	0.063	22.68
0.5, 0.5	-0.53	94.8	1	0.064	39.86
0.4, 0.6	-3.19	93.4	1	0.063	22.08
0.3, 0.7	-10.21	89	1	0.059	6.29
0.2, 0.8	-19.66	77.8	1	0.053	1.0
0.1, 0.9	-29.83	51.9	1	0.048	0.3
DIC	-0.67	94.7	1	0.064	
Swapping between Groups of 4					
0.9, 0.1	-33.83	45.5	0.998	0.048	0.1
0.8, 0.2	-22.40	71.6	0.997	0.053	0.4
0.7, 0.3	-12.05	89.4	0.999	0.059	6.0
0.6, 0.4	-4.64	94.1	0.999	0.062	23.5
0.5, 0.5	-1.97	95.4	1	0.064	38.9
0.4, 0.6	-4.75	93.5	1	0.062	24.0
0.3, 0.7	-12.24	87.5	1	0.059	6.5
0.2, 0.8	-22.63	71.4	1	0.053	0.5
0.1, 0.9	-34.10	44.6	1	0.048	0.1
DIC	-0.59	95.2	1	0.063	

The Type A, N=5760, N_p^I=240, 24 cases per interviewer at both the current and previous wave, W_{ij}=(0.5, 0.5), σ_u^2=0.3 and π=0.8 factor values are used to analyse the effect of limited pairings on the properties of the estimator

and test statistic, and the DIC reliability. The results presented in Table 22 show that while the properties are not affected by the amount of mixing in interviewer pairings for comparable Type A scenarios, the DIC does perform slightly better when the cases are switched randomly across all possible interviewer compared to the two more restrictive scenarios. There is no difference in the results for the reliability of the DIC between the scenario with pair swapping and the scenario for groups of four swapping.

In reality the comparison of Type E to Type C and Type F to Type D change profiles also gives an indication of the impact of limited pairings on the model and DIC performance on different change profile types. Similar results to the Type A example presented above are obtained. No evidence of worse properties for Type E and F compared to Type C and D is found. The DIC reliability measure performs worse for the change profiles with limited pairings which allocate a whole caseload of a dropped interviewer to another interviewer (Type E and Type F) compared to their equivalent scenarios (Type C and Type D) which distribute cases from dropped interviewers randomly across all available interviewers.

IV.5.7. Implications of Running 2-Level Models for Multiple Membership Data

This section will explore the properties of the estimator and test statistic for multiple membership data run using a simple 2-level model which simply includes a random effect for the current interviewer. Moreover, the DIC reliability measure is computed again this time identifying the model with the lowest DIC from 10 possibilities – the nine multiple membership models and the 2-level model. The proportion of times the 2-level model will correspond to the lowest DIC gives a measure of how often multiple membership data can be wrongly modelled as a simple hierarchical 2-level structure when basing the model choice on an empirical assessment method (DIC).

In Table 24, for the purely hierarchical models, higher negative bias and lower confidence interval coverage rates are observed for scenarios with a higher percentage of cases with multiple memberships. This result is

consistent with the work by Chung and Beretvas (2012). The authors find underestimated higher-level variance when modelling multiple membership data using simple 2-level models. This decrease is more pronounced for scenarios with a higher proportion of cases experiencing membership change.

Table IV.24: Properties of the Estimator and Test Statistic and Percentage Of Simulations which Corresponds with Lowest DIC for the 2-Level Models

Type	% Change	W_{ij}	N	Bias	CI Cov.	Power	S.E.	% Lowest DIC
A	8	0.5, 0.5	2880	-5.6	91.0	1.00	0.078	9
	8	0.9, 0.1	2880	2.0	95.6	1	0.081	9.5
	50	0.5, 0.5	2880	-47.4	36.6	0.89	0.059	8.5
	50	0.9, 0.1	2880	-10.0	88.4	1.00	0.076	13.5
	8	0.5, 0.5	5760	-7.9	88.8	1	0.053	8
	8	0.9, 0.1	5760	-0.4	94.4	1	0.055	11.6
	50	0.5, 0.5	5760	-45.0	14.3	0.99	0.041	7.6
	50	0.9, 0.1	5760	-8.8	88.2	1	0.053	14.2
B	8	0.5, 0.5	2880	-9.3	89.3	1.00	0.076	9
	8	0.9, 0.1	2880	-0.9	94.6	1.00	0.080	10.5
	50	0.5, 0.5	2880	-47.1	38.0	0.88	0.059	9.9
	50	0.9, 0.1	2880	-8.8	87.9	1.00	0.076	10.8
	8	0.5, 0.5	5760	-7.3	89.0	1	0.053	8.9
	8	0.9, 0.1	5760	-1.3	93.6	1	0.055	8.9
	50	0.5, 0.5	5760	-45.3	14.6	1.00	0.041	9.4
	50	0.9, 0.1	5760	-8.7	90.0	1	0.053	13.9
C	50	0.5, 0.5	5760	-39.3	26.9	1.00	0.043	6.3
	50	0.9, 0.1	5760	-9.2	88.7	1	0.053	14.7
D	50	0.5, 0.5	5760	-38.3	29.4	1.00	0.043	9.6
	50	0.9, 0.1	5760	-9.4	88.1	1	0.053	13.8
E	50	0.5, 0.5	5760	-25.1	60.7	1	0.048	9
	50	0.9, 0.1	5760	-7.8	89.1	1	0.053	11.9
F	50	0.5, 0.5	5760	-25.9	57.8	1	0.047	8.4
	50	0.9, 0.1	5760	-8.5	89.1	1	0.053	8.9

The DIC's ability to detect the MM structure dos not seem to be affected by the percentage of cases with multiple memberships. Poorer properties are observed for scenarios with W_{ij} =(0.5, 0.5) compared to W_{ij} =(0.9, 0.1) scenarios. This result makes intuitive sense, as W_{ij}=(0.9, 0.1) data is more congruent with a 2-level structure (equivalent to W_{ij}=1, 0) than the W_{ij}=(0.5, 0.5) data. By the same logic, the percentage of 2-level models with the lowest DIC is generally higher for W_{ij} =(0.9, 0.1) scenarios than W_{ij} =(0.5, 0.5) scenarios. When the sample size N is halved (considered for Type A and Type B scenarios), lower power and surprisingly higher confidence interval coverage are obtained for 50% change, W_{ij}=(0.5, 0.5) scenarios, as well as consistently higher standard errors. A change in N does not seem to affect bias or the percentage of times the lowest DIC is obtained for the 2-level model. Some variation in the percentage of times the 2-level model corresponds with the lowest DIC is observed across change profile types.

With the DIC selection method, the 2-level model is selected as the best fitting model over nine other competing multiple membership models only a maximum of 15% in the scenarios considered. This result indicates that the DIC is able to identify that multiple membership models fit multiple membership data better than 2-level models in the majority of cases for the scenario considered.

IV.6. Discussion

This paper investigated the properties of the variance estimator and the test statistic for multiple membership models with different weight specifications. The different models include all possible theoretically-based weights, with a one decimal place precision level, and models based on the weights identified as giving best fit by the DIC measure. Moreover, the reliability of the DIC in identifying the true multiple membership weights has also been examined. These measures have been explored for data with different true multiple

membership weights (W_{ij}), proportion of cases associated with multiple memberships (percentage change), membership profiles (change profile types) and total sample size (N).

As expected, the results show optimal properties for models specifying the correct model weights w_{ij}. These properties generally do not vary across different factor values for the models including the correct weights w_{ij}. One exception is the increase in standard errors with smaller N and with higher proportions of multiple membership cases. In comparison, models with misspecified w_{ij} obtain less than optimal, and at times alarmingly bad results. DIC-based weights models obtain good results overall, sometimes reaching values equivalent to the models including the correct weights. The different factors interact with each other in a complex way to influence the properties of the estimator and test statistic, and the DIC reliability. Also, there often seems to be a threshold beyond which an increase in a specific factor value influences these measures. This threshold varies for different combinations of simulation factor values and for the different measures. Though some general trends can be observed, this study highlights the need for each particular application (with its particular data structure) to be considered individually, to inform decisions on either data collection design or inference.

A higher proportion of cases experiencing multiple memberships, represented by the factor percentage change, generally leads to worse estimator properties and power of the Wald test for models including incorrect w_{ij}. Estimator bias is highly sensitive to variation in the percentage change factor. Drastic bias increases with higher percentage change values are observed for all percentage change values considered, with a greater increase observed for the worst specified models. Lower power with higher MM proportions is only observed from very high percentage change values (92%) for N=5760 and from medium percentage changes (50%) for N=2880 scenarios for models with misspecified weights ($w_{ij} \neq W_{ij}$). Confidence interval coverage also shows a systematic decrease for the whole spectrum of percentage change values considered for misspecified w_{ij} models, therefore resulting in greater underestimation for incorrect models. Standard errors

show an increase for models with correct weights (w_{ij}=W_{ij}) and a decrease for models with misspecified weights (w_{ij}≠W_{ij}) with percentage change increases. These patterns observed for the models with theoretical w_{ij} do not apply for DIC-based models. Variations in percentage change have no impact on the power of the Wald test, as power is close to 1 for all scenarios, and no consistent effect on estimator bias and confidence interval coverage for DIC-based weights models. On the other hand, a clear trend of larger standard errors with increasing percentage change is observed. This result is similar to the trend observed for theoretical weights models with correct w_{ij}. The DIC reliability measure shows a clear trend of better results for data with higher proportions of MM cases.

Halving N, and consequently the number of interviewers, results in lower power for the 50% change scenarios (no change is observed for the 8% change scenarios) for misspecified w_{ij} models, slightly lower CI coverage for models with correct w_{ij} and higher CI coverage for models with incorrect w_{ij} for the 50% change scenarios and larger standard errors for all models with theoretical weights. No consistent pattern in the estimator bias across N values is observed. For the DIC-based weights models no relationship between N and power, CI coverage and bias is observed. On the other hand, larger standard errors for smaller sample size scenarios are obtained for DIC-based w_{ij} models. Decreases in N also impact negatively the DIC reliability measure. This relationship is mediated by other factors, including W_{ij} and the profile change type.

The effect of W_{ij} on the properties varies for different factor combinations. However, some general trends can be identified, though exceptions to this trend are noticeable for certain scenarios. One common feature across different change profiles is the general shape of the distribution of the estimator bias across different theoretical weights models. The bias for W_{ij}=(0.5, 0.5) scenarios are somewhat symmetrical around the correct model [(w_{ij}=0.5, 0.5)]. W_{ij}=(0.9, 0.1) scenarios generally show low positive bias for the correct model [w_{ij} =(0.9, 0.1)], which increases in magnitude for subsequent models and then eventually turns negative and increases for

models with weights closer to (0.1, 0.9). For the CI coverage a symmetry around the w_{ij}=(0.5, 0.5) model is observed for the equally distributed W_{ij}, while a stable and then decreasing trend from the w_{ij}=(0.9, 0.1) model to the w_{ij}=(0.1, 0.9) model is observed for the unequally distributed W_{ij}.

In the case of the power of the Wald test optimal values are obtained for N=5760 scenarios. Therefore, for large sample sizes no difference across W_{ij} scenarios is observed. Smaller sample sizes are considered for Type A and Type B change profiles. For 50% change, N=2880 scenarios higher power is obtained for Type A in comparison to lower power for Type B for W_{ij}=(0.5, 0.5) scenarios. Distinct patterns are present for the size of the standard errors by different W_{ij}. For W_{ij}=(0.9, 0.1) scenarios the w_{ij}=(0.9, 0.1) model has higher standard errors than the model with the worst specified weights, but lower standard errors than the w_{ij}=(0.5, 0.5) model. For W_{ij}=(0.5, 0.5) scenarios the distribution across the models with theoretically-based weights is symmetrical, with the model specifying the w_{ij}=(0.5, 0.5) model having the highest standard errors. For the DIC-based weights models the results show a clear trend of higher bias for W_{ij}=(0.9, 0.1) scenarios compared to W_{ij}=(0.5, 0.5) scenarios for the 50% change scenarios. The higher biases for more unequally distributed real weights data is not observed for scenarios with a low percentage of cases with multiple memberships. No change across the W_{ij} values for power, confidence interval coverage and standard errors are observed. The DIC is better able to detect the model with the correct weights for more unequal W_{ij} data, and this difference is more pronounced for data with lower percentage change.

The properties of the estimator and test statistic and the DIC reliability, as well as their relationship with other factors, vary by change profile type. Some noticeable general trends are the observation of skewness for the distribution of bias and CI coverage for models with theoretically-based weights for Type B, Type C and Type E. These change profile types vary in their number of interviewers and caseload across the two waves, unlike the other change profile types. Higher CI coverage and less underestimated standard errors are obtained for W_{ij} =(0.9, 0.1) for the models with misspecified

theoretically-based weights for the change profiles with a higher number of interviewers (Type B, Type D, Type F) compared to the other change profiles (Type A, Type D, Type E). Possibly the influence of a higher number of clusters on these properties is restricted to the W_{ij}=(0.9, 0.1) scenarios because a higher level of misspecification is possible for unequally distributed real weights data, and the influence of the number of clusters is only observed for very badly misspecified models. The impact of N on power seems to change by profile type, as outlined above. An important result is that for models including the correct (and adjacent) theoretical weights and the models including weights based on DIC there is no substantial difference in the properties of the estimator and Wald test across different change profile types. The DIC reliability measure also varies by change profile type, showing the best results for Type A, and the worst for Type F. On closer inspection, for the Type F profile the mean DIC values are almost equal for the different models with theoretical weights. This lower DIC reliability does not result in worse properties of estimators and test statistics for the model based on the DIC, expect for a 5% lower confidence coverage rate in comparison to the model with correct w_{ij} for W_{ij}=(0.5, 0.5).

The study has also investigated the ability of the DIC to distinguish multiple membership data from purely hierarchical data. The results are encouraging, showing that the lowest DIC value corresponds to the 2-level model (ignoring the multiple membership structure) less than 10% of times for W_{ij} =(0.5, 0.5) scenarios, and less than 15% of times for W_{ij} =(0.9, 0.1) scenarios. Specifying a purely hierarchical model for multiple membership data results in high bias, low confidence interval coverage, low power of the Wald test and underestimated standard errors, especially for scenarios with a high percentage of cases with multiple memberships and with equal real weights [W_{ij}=(0.5, 0.5)].

IV.7. Conclusion

The results would suggest that before deciding on the method to choose the weights the characteristics of the data should be noted. For example, when the data includes a low percentage of multiple memberships, empirical assessment methods, such as the DIC, may not suggest the true multiple membership weights, and it may be better to base the choice of weights on some theoretical argument if the researcher is particularly interested in the substantive interpretation of the weights. However, all nine models with the different theoretical weights obtain good properties of the estimator and test statistic for scenarios with a low proportion of multiple membership cases. Therefore, to the extent that the researcher is only interested in the variance estimate, any reasonable weighting scheme can be applied when only a low percentage of cases are associated with more than one higher-level unit. What constitutes a low percentage will change depending on the other factor values, such as N and the change profile type. One needs to be careful when interpreting the substantive meaning of the model weights as the frequency with which the DIC is able to detect the correct model weights can be low. Rather than speaking of exact proportions for the higher-level influences it may be best to refer more loosely to the variance apportionment. The simulation study has also highlighted the effectiveness of the DIC in distinguishing between purely hierarchical and multiple membership data.

The results show that despite inaccurate multiple membership weights, models including DIC-based weights result in good estimator properties and power of the Wald test. To the extent that the multiple membership weights are not of substantive interest, it may be best to always choose weights based on the DIC.

When analysing data with a multiple membership structure the true apportionment of the total variance across the two membership classification structures will be unknown. The DIC will be used to identify the best fitting model amongst various models specifying different model weights. The model weights may be interpreted as the weights signifying the true variance proportions. If the model with the lowest DIC does not correspond with the

model which best estimates the relative influence of the multiple memberships giving a substantive interpretation to the model weights will be erroneous. This study shows that for certain data structures the DIC fails to identify the true MM weights (that is, the weights used to simulate the data), but consistently results in good properties for the variance estimators and the test statistic. Consequently, it may be best to think of the model weights simply as parameters which optimise a certain information criterion, which may or may not represent the true influence apportionment of the two higher-level classification structures.

V. Conclusion

This study focuses on the use of cross-classified and multiple membership multilevel logistic models for the analysis of interviewer effects on wave nonresponse in longitudinal surveys. Papers 1, 2 and 3 all demonstrate how cross-classified and multiple membership multilevel logistic models provide a flex class of models for the analysis of interview effects on nonresponse in longitudinal surveys. Paper 1 introduces the mathematical details of and the assumptions underlying these models. A dataset from the UK Family and Children Study is used to illustrate the proposed methods and to investigate substantive questions regarding interviewer characteristics which explain the variation across interviewers in individual-level nonresponse. Paper 2 investigates the properties of estimators and test statistics for cross-classified multilevel models accounting for area and interviewer effects on nonresponse for data with varying degrees of interpenetration between the two classification structures. Paper 3 investigates the properties of estimators and test statistics for multiple membership multilevel models accounting for distinct interviewers allocated in the current and previous wave for data with different interviewer allocation profiles. This section provides a summary of the main findings from the three papers; in addition, any limitations are highlighted and recommendations for future work are proposed.

In Paper 1 both cross-classified and multiple membership specifications are considered to account for the effect of various wave interviewers allocated to a particular case. The inclusion of the previous and current wave interviewer effects as independent effects in a cross-classified model – as implemented in Pickery et al. (2001) – led to unstable results in the Paper 1 application, suggesting model misspecification. The assumption of independent effects certainly does not apply for the 32% of cases with interview continuity. For these cases there may either be a correlated differential effect across waves or alternatively a single constant effect (suggesting a multiple membership structure). This problem with specifying two wave interviewer effects as cross-

classified has also been highlighted by Lynn et al. (2013). In the study by Pickery et al. (2001) the second wave interviewer becomes non-significant when both second and first wave interviewer effects are controlled for. They recognise that the inclusion of both previous and current wave interviewer effects as independent effects is problematic due to the correlation between these two random effect terms. However the authors believe that this correlation does not arise from the partial overlap of the two wave interviewers (owing to cases with interviewer continuity), since this overlap information is lacking in the model. Instead they attribute this correlation to the geographical setup of the survey, with case assignments being stable across waves in terms of geographical area.

For the Paper 1 application, the multiple membership model does not provide a better model fit than a 2-level model accounting only for the current interviewer when measurable interviewer characteristics are not accounted for. Among competing multiple membership models specifying different weights for the interviewer change cases – varying from weights of (0.9, 0.1) to weights of (0.1, 0.9), the model allocating the highest weight of 0.9 to the current wave interviewer fits best. This suggests that for later waves of a longitudinal study, the current interviewer has the greatest impact on the response outcome for the current wave. The interviewer variance estimates for the multiple membership models with different weight specifications (which range from 0.250 to 0.291) vary slightly from the 2-level model, accounting only for the wave 8 interviewer (0.278). If the 2-level model is considered to be the true model, the deviation from this 2-level model estimate obtained from modelling the data using a multiple membership model ranges from -8.4% to 6.6%. However the best fitting multiple membership model, including weights of (0.9, 0.1) for change cases, obtains practically the same estimate as the 2-level model (0.282). Stable variance estimates across different weight specifications have been documented in other applied studies in different substantive areas using multiple membership models (Fielding, 2002; Fielding & Yang, 2005; Goldstein, 2011b). These studies do not indicate what alternative weighting profiles were considered and the estimates obtained across different models with different weights. It could be that this reported stability across weighting profiles is due to the fact that these authors do not consider weighting profiles

which deviate greatly from the correct weights. In the Paper 1 application the estimates obtained across all possible weighting profiles (with a precision of 0.1) cannot be said to be stable as the highest and lowest estimate value vary by around 15%.

The application of multiple membership models to interviewer effects data is relatively new. The only other research using this model specification for the analysis of nonresponse is a manuscript by Lynn et al. (2013), which finds a non-significant random intercept interviewer effect. In the Lynn et al. (2013) paper a 2-level model accounting for the current interviewer, a 2-level model accounting for the previous interviewer, a multiple membership model with equal weights, and a model with only a fixed-effect intercept all obtain approximately the same DIC value. However the paper does find a differential random effect in the impact of respondent age on the propensity to refuse participation by interviewer combination (random slope at the interviewer-level for respondent age). In spite of limited evidence of the presence of a multiple membership structure, this model specification for the analysis of interviewer effects on wave nonresponse should not be discarded. The fitting of this model to various other datasets analysing nonresponse at various waves is encouraged to corroborate evidence on the influence of multiple distinct interviewers from different waves on nonresponse at a particular wave. It would be interesting to consider whether different results are obtained for earlier waves in comparison to later waves when the remaining sample includes people with a high commitment towards the survey.

For the Paper 1 application, the 2-level model accounting for the wave 8 interviewer had a better model fit and accounted for a larger proportion of the variance in individual nonresponse than the 2-level model accounting for the wave 7 interviewer. The opposite is true in Pickery and Loosveldt (2001), where these same model specifications showed that the first wave interviewer had a stronger influence on wave 2 nonresponse than the second wave interviewer. Moreover, in a cross-classified model including both interviewer effects the wave 2 interviewer random effect became non-significant at the 5% level. This discrepancy in results may be due to the fact that these two studies are looking at different phases of a longitudinal study.

A cross-classified model is advocated for distinguishing area and interviewer effects in the case of partial interpenetration, which is sometimes present in surveys. Area effects are not significant after controlling for interviewer and household level effects in a cross-classified model in the Paper 1 application. This result is similar to findings by Campanelli and O'Muircheartaigh (1999) and Durrant et al. (2010). The results from Paper 2 suggest that limited interpenetration is sufficient to correctly disentangle the two random effects. Consequently the non-significance of the cross-classified area effect in comparison to the significant area effect in a 2-level model suggests that area effects are simply aggregated interviewer effects. Reference is made to the possibility that the primary sampling unit does not match the spatial divisions which are related to nonresponse. This area classification is used in Paper 1 and also in other studies analysing nonresponse for data with a multistage cluster sample design (Durrant et al., 2010) as well as data from a quasi-randomised design (Campanelli & O'Muircheartaigh, 1999). Possible theoretical arguments for area effects on nonresponse include 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). Future research may investigate the sensitivity of the area parameter estimate to changes in the area classification system, with different area boundaries mapping to physical, social and cultural spatial divisions being considered.

The substantive findings from Paper 1 confirm that interviewer experience, grade and continuity are significant predictors of nonresponse. The results suggest improved response rates for interviewers who commit most of their paid working hours to undertaking interviewing work for various survey agencies, and who have better job tenure and are focused on face-to-face mode. The study does not provide any clear or coherent evidence of the role of interviewer attitudes on respondent persuasion, personality traits and skills on respondent refusal. Similarly no evidence of the benefit of matching the individual and the interviewer in terms of demographic or socio-economic characteristics was found.

The simulations in Paper 2 and Paper 3 offer new insight into the performance of the advanced multilevel models for realistic survey design conditions. Paper 2 is the first work investigating these properties for cross-classified models, and similarly Paper 3 is the first study investigating these properties for multiple membership models when the true multiple membership weights are unknown, as would be the case in a real life situation. The two simulation studies identify trends in the properties of the estimators and test statistic across changes in simulation factors. Additionally, in the case of Paper 3, the study also indicates how reliable the DIC is in detecting the real weights and how this varies for different interviewer allocation schemes. It is acknowledged that the results from Paper 2 and Paper 3 are restricted to the factor values chosen and the scenarios considered. The results cannot be extrapolated to very different survey design conditions with any certainty. Moreover, a potential restriction of Paper 2 and Paper 3 is the limited number of scenarios considered in view of the time constraints in running the simulations. Due to the small number of scenarios considered, significance tests were not run to identify the factors which are significantly associated with the properties examined.

Paper 2 indicates that, as expected, purely hierarchical data, represented by CASE 1 allocation scheme scenarios, is subject to substantial biases, bigger 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. Limited interviewer dispersion (of around 3 areas per interviewer for medium or large sample sizes, N=2880 or N=5760) provides sufficient interpenetration for good properties. Further dispersion yields only very small or negligible gains in the properties. Interviewer dispersion 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.

Paper 3 shows optimal properties for models specifying the correct or adjacent model weights, which are relatively stable across different factor conditions. The only exception is the higher standard errors for scenarios with

smaller sample sizes. Poor estimator and test statistic properties are obtained for models including very badly misspecified weights. On the other hand, models with weights based on the DIC generally obtain good properties which are similar to the values obtained for the models specifying the correct weights in most scenarios. The results indicate that the DIC does not offer a very precise measure for choosing the correct model weights. However, if selecting either the correct weights or the adjacent weights is deemed acceptable, the results are more encouraging. Consequently the substantive interpretation of the weights should be carried out with caution. The paper provides evidence for the reliability of the DIC to detect a multiple membership structure, in comparison to a 2-level model accounting only for the current interviewer which obtains the lowest DIC only a maximum of 15% of times.

The papers provide a good starting point for the analysis of the performance of these models under different scenarios, but they definitely do not present a conclusive and comprehensive overview. 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. The simulations always consider small to medium variance values, and the simulation papers deliberately focus on variance values which are considered to correspond to realistic small to medium variance partitioning coefficient values, signifying substantial higher-level effects. Therefore other studies investigating the properties of the estimators and test statistic for very small higher-level variances should be carried out.

The two simulation papers considered the properties of variance estimators only. The data was generated from models including an overall intercept and the random effects. No explanatory variables were 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.

Paper 2 considers only scenarios where the interviewers are allocated cases from different areas randomly and Paper 3 considers only scenarios where the interviewer effects for the multiple membership structure are independent and identically distributed. Therefore the results obtained in the simulations apply only to similar situations for which these assumptions hold. Situations of non-random allocations and correlated effects are not being investigated.

This work created the procedure and the R and STATA programming code – included in the Appendices – that can be used independently of this research project to investigate the performance of these multilevel 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.

A further simulation exercise has been run for this concluding chapter. The aim of this simulation exercise is to identify how other multilevel model specifications perform for 2-level data and how reliable the DIC is in detecting the correct hierarchical structure. One thousand datasets were simulated from the final random effects specification chosen for Paper 1: the 2-level model accounting for the wave 8 interviewer with a higher-level variance equal to 0.273 and an overall probability of 0.91. The sample size and the data structure (the interviewer allocations and area provenance for each individual case) of the Family and Children Study dataset are used for the simulation. These datasets were then modelled using the correct model (2-level model including the wave 8 interviewer random effect) as well as alternative models, these being multiple membership models with different weights, 2-level model including the area interviewer random effect, and cross-classified models.

Table V.1: Frequency Distribution of Models with Lowest DIC

Model	Percentage
2-level Model; Interviewer Wave8	42.0
Cross-classified Model; Interviewer Wave 8 & Primary Sampling Unit	12.4
Multiple Membership Model; Weights (0.9 0.1)	12.3
Multiple Membership Model; Weights (0.8 0.2)	9.8
Cross-classified Model; Interviewer Wave 8 & Interviewer Wave 7	7.6
Multiple Membership Model; Weights (0.7, 0.3)	6.2
Multiple Membership Model; Weights (0.6, 0.4)	3.4
Multiple Membership Model; Weights (0.5, 0.5)	2.0
Multiple Membership Model; Weights (0.1, 0.9)	1.6
Multiple Membership Model; Weights (0.4, 0.6)	1.0
Multiple Membership Model; Weights (0.2, 0.8)	0.9
Multiple Membership Model; Weights (0.3, 0.7)	0.5
2-level Model; Primary Sampling Unit	0.4
Total	100

Table 1 shows the percentage of times the lowest DIC value for the 1000 datasets corresponds to a specific model. Interestingly, the correct model is only detected 42% of the times. The next two most frequent models with the lowest DIC values are the multiple membership models attributing weights of (0.9, 0.1) to the wave 8 and wave 7 interviewers respectively, and the cross-classified model controlling for the wave 8 interviewer and the primary sampling unit. Table 2 presents the properties of the interviewer variance estimator, the power of the Wald test, and the mean DIC for these three models. The multiple membership model with (0.9, 0.1) weights shows a slight overestimation of the interviewer variance. On the other hand, the cross-classified model also controlling for area effects shows a slight underestimation of the interviewer variance. Similar results are observed in Paper 1, where the multiple membership model with (0.9, 0.1) weights obtains a higher estimate, while the cross-classified model obtains a lower estimate, than the 2-level model. For this final simulation, other than this slight bias observed for the two misspecified models, the other properties are close to optimal for all three models. The area effect in the cross-classified model is non-significant at the 10% significance level and the mean DIC values are very similar across the three models, as per the results in Paper 1. These results

support the final conclusion in Paper 1 which stated that it was sufficient to only control for the wave 8 interviewer effect for this particular dataset.

Table V.22: Properties of the Interviewer Variance Estimator and Test Statistic for Different Models

Model	Wald CI Coverage	Power Wald Test	Percentage Relative Bias	Standard Error	Mean DIC
2-level Model; Interviewer Wave 8	92.0	0.98	-0.45	0.078	3617.9
Cross-classified Model; Interviewer Wave 8 & Primary Sampling Unit	90.6	0.95	-6.02	0.081	3618.4
Multiple Membership Model; Weights (0.9 0.1)	92.6	0.98	4.93	0.083	3618.2

Therefore, the results of this final simulation show that the DIC is able to identify a good model – either the correct model or an alternative model which is close to the correct model – with a high frequency. This corroborates evidence from other studies examining the performance of the DIC in identifying the correct model (Berg et al., 2004; Kizilkaya & Tempelman, 2003; Ward, 2008; Wilberg & Bence, 2008; Zhu & Carlin, 2000). Moreover, the alternative models show good properties of the estimator and test statistic. A possible extension of this analysis would be to simulate from the other possible true models fitted in Paper 1, including cross-classified models, multiple membership models, and the 2-level model accounting for the PSU. It would be interesting to assess frequency with which the DIC identifies the correct model and the adequacy of the models with next highest frequency of lowest DIC values. The properties of these alternative models would indicate their adequacy.

Further research into the optimal random effects specification for the analysis of interviewer effects on nonresponse in a longitudinal study is required. Cross-classified models assume that the interviewer effect is wave-

specific and introduce a separate independent effect for each wave, but fail to account for the correlation of different wave-interviewer effects when the same interviewer is maintained across waves. Multiple membership models incorporate the effect of each distinct interviewer associated with a particular case, but make no attempt to distinguish between the differential effects of the same interviewer across waves. An extension which could be attempted in future entails the inclusion of a covariance term for the cross-classified random effect specification. However, just as the independence assumption of the cross-classified model attempted in the first paper was assessed to be erroneous as it does not take into consideration interviewer continuity cases, the specification including a covariance term for the two higher-level variances would not appropriately model the interviewer effects for interviewer change cases.

At present no model can be identified to allow the specification of two separate terms – specified as correlated for cases retaining the same interviewer and as independent for cases experiencing an interviewer change – for the two different interviewer-wave effects for all cases. Such a model would enable a distinction between the effects of each interviewer at either wave while recognising cases with interviewer continuity. Moreover, this model would allow the relative influence of each wave interviewer for all cases – identified as one of the aims of this study – to be estimated. Such a model has been identified as an area of potential new methodological development.

VI. Appendices

VI.1. Appendix A – Descriptive Statistics for Complete and Restricted Datasets

Frequency Distribution of the Response Outcome by Categorical Variables for the Individual-Level Data

Category	Statistic	Productive Interview	Refusals	Total
Government Office Region (5932 cases)				
North East	Count	349	28	377
	%	92.6%	7.4%	100%
North West & Merseyside	Count	635	74	709
	%	89.6%	10.4%	100%
Yorkshire & Humber	Count	547	44	591
	%	92.6%	7.4%	100%
East Midlands	Count	470	33	503
	%	93.4%	6.6%	100%
West Midlands	Count	587	49	636
	%	92.3%	7.7%	100%
South West	Count	535	47	582
	%	91.9%	8.1%	100%
Eastern	Count	439	40	479
	%	91.6%	8.4%	100%
London	Count	403	52	455
	%	88.6%	11.4%	100%
South East	Count	633	69	702
	%	90.2%	9.8%	100%
Wales	Count	360	35	395
	%	91.1%	8.9%	100%
Scotland	Count	473	30	503
	%	94.0%	6.0%	100%
Government Office Region (7089 cases)				
North East	Count	393	34	427
	%	92.0%	8.0%	100%
North West & Merseyside	Count	690	78	768
	%	89.8%	10.2%	100%
Yorkshire & Humber	Count	642	49	691
	%	92.9%	7.1%	100%
East Midlands	Count	555	48	603
	%	92.0%	8.0%	100%
West Midlands	Count	698	58	756
	%	92.3%	7.7%	100%

South West	Count	581	55	636
	%	91.4%	8.6%	100%
Eastern	Count	526	54	580
	%	90.7%	9.3%	100%
London	Count	526	72	598
	%	88.0%	12.0%	100%
South East	Count	875	85	960
	%	91.1%	8.9%	100%
Wales	Count	386	40	426
	%	90.6%	9.4%	100%
Scotland	Count	599	45	644
	%	93.0%	7.0%	100%
London Indicator (5932 cases)				
Not London	Count	5028	449	5477
	%	91.8%	8.2%	100%
London	Count	403	52	455
	%	88.6%	11.4%	100%
London Indicator (7089 cases)				
Not London	Count	5945	546	6491
	%	91.6%	8.4%	100%
London	Count	526	72	598
	%	88.0%	12.0%	100%
How much of a problem are vandalism, graffiti and other deliberate damage to property or vehicles? (5932 cases)				
Very big problem/Fairly big problem	Count	945	77	1022
	%	92.5%	7.5%	100%
Not a very big problem	Count	1747	142	1889
	%	92.5%	7.5%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	2739	282	3021
	%	90.7%	9.3%	100%
How much of a problem are vandalism, graffiti and other deliberate damage to property or vehicles? (7089 cases)				
Very big problem/Fairly big problem	Count	1118	91	1209
	%	92.5%	7.5%	100%
Not a very big problem	Count	2097	186	2283
	%	91.9%	8.1%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	3256	341	3597
	%	90.5%	9.5%	100%
How much of a problem is rubbish or litter lying around?(5932 cases)				
Very big problem/Fairly big problem	Count	988	89	1077
	%	91.7%	8.3%	100%
Not a very big problem	Count	1943	158	2101
	%	92.5%	7.5%	100%
Not a problem at all/It happens but it is not a	Count	2500	254	2754

problem/Don't know	%	90.8%	9.2%	100%
How much of a problem is rubbish or litter lying around?(7089 cases)				
Very big problem/Fairly big problem	Count	1141	106	1247
	%	91.5%	8.5%	100%
Not a very big problem	Count	2317	198	2515
	%	92.1%	7.9%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	3013	314	3327
	%	90.6%	9.4%	100%
How much of a problem are teenagers hanging around on the street? (5932 cases)				
Very big problem/Fairly big problem	Count	1331	107	1438
	%	92.6%	7.4%	100%
Not a very big problem	Count	1518	120	1638
	%	92.7%	7.3%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	2582	274	2856
	%	90.4%	9.6%	100%
How much of a problem are teenagers hanging around on the street? (7089 cases)				
Very big problem/Fairly big problem	Count	1567	132	1699
	%	92.2%	7.8%	100%
Not a very big problem	Count	1794	157	1951
	%	92.0%	8.0%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	3110	329	3439
	%	90.4%	9.6%	100%
How much of a problem are troublesome neighbours? (5932 cases)				
Very big problem/Fairly big problem	Count	386	35	421
	%	91.7%	8.3%	100%
Not a very big problem	Count	682	57	739
	%	92.3%	7.7%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	4363	409	4772
	%	91.4%	8.6%	100%
How much of a problem are troublesome neighbours? (7089 cases)				
Very big problem/Fairly big problem	Count	469	46	515
	%	91.1%	8.9%	100%
Not a very big problem	Count	833	71	904
	%	92.1%	7.9%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	5169	501	5670
	%	91.2%	8.8%	100%
How much of a problem is people being attacked or harassed because of their skin colour, religion or ethnic origin? (5932 cases)				
Very big problem/Fairly big problem	Count	167	19	186
	%	89.8%	10.2%	100%
Not a very big problem	Count	546	34	580
	%	94.1%	5.9%	100%
Not a problem at all/It happens but it is not a	Count	4718	448	5166

problem/Don't know	%	91.3%	8.7%	100%
How much of a problem is people being attacked or harassed because of their skin colour, religion or ethnic origin? (7089 cases)				
Very big problem/Fairly big problem	Count	213	23	236
	%	90.3%	9.7%	100%
Not a very big problem	Count	648	45	693
	%	93.5%	6.5%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	5610	550	6160
	%	91.1%	8.9%	100%
How much of a problem are people being drunk or rowdy in public places? (5932 cases)				
Very big problem/Fairly big problem	Count	666	56	722
	%	92.2%	7.8%	100%
Not a very big problem	Count	1207	103	1310
	%	92.1%	7.9%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	3558	342	3900
	%	91.2%	8.8%	100%
How much of a problem are people being drunk or rowdy in public places? (7089 cases)				
Very big problem/Fairly big problem	Count	789	67	856
	%	92.2%	7.8%	100%
Not a very big problem	Count	1457	129	1586
	%	91.9%	8.1%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	4225	422	4647
	%	90.9%	9.1%	100%
How much of a problem are people using or dealing drugs? (5932 cases)				
Very big problem/Fairly big problem	Count	823	67	890
	%	92.5%	7.5%	100%
Not a very big problem	Count	776	64	840
	%	92.4%	7.6%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	3832	370	4202
	%	91.2%	8.8%	100%
How much of a problem are people using or dealing drugs? (7089 cases)				
Very big problem/Fairly big problem	Count	977	86	1063
	%	91.9%	8.1%	100%
Not a very big problem	Count	924	78	1002
	%	92.2%	7.8%	100%
Not a problem at all/It happens but it is not a problem/Don't know	Count	4570	454	5024
	%	91.0%	9.0%	100%
Likelihood that purse/wallet is returned to you if found in the street by someone living in your neighbourhood(5932 cases)				
...very likely, quite likely,	Count	3052	292	3344
	%	91.3%	8.7%	100%
not very likely,	Count	1281	109	1390
	%	92.2%	7.8%	100%
or not at all likely? Don't Know	Count	1098	100	1198

	%	91.7%	8.3%	100%
Likelihood that purse/wallet is returned to you if found in the street by someone living in your neighbourhood(7089 cases)				
...very likely, quite likely,	Count	3591	350	3941
	%	91.1%	8.9%	100%
not very likely,	Count	1542	142	1684
	%	91.6%	8.4%	100%
or not at all likely? Don't Know	Count	1338	126	1464
	%	91.4%	8.6%	100%
Wave 8 Interviewer Grade/ Experience (5932 cases)				
A; B	Count	354	42	396
	%	89.4%	10.6%	100%
C, 0-4 yrs	Count	1919	229	2148
	%	89.3%	10.7%	100%
C, 5 yrs+	Count	1049	68	1117
	%	93.9%	6.1%	100%
D, 0-4 yrs	Count	472	52	524
	%	90.1%	9.9%	100%
D, 5 yrs+	Count	533	48	581
	%	91.7%	8.3%	100%
S, 0-4 yrs	Count	335	30	365
	%	91.8%	8.2%	100%
Grade R; S, 5 yrs+; T	Count	769	32	801
	%	96.0%	4.0%	100%
Wave 8 Interviewer Grade/ Experience (7089 cases)				
A; B	Count	553	73	626
	%	88.3%	11.7%	100%
C, 0-4 yrs	Count	2356	287	2643
	%	89.1%	10.9%	100%
C, 5 yrs+	Count	1249	78	1327
	%	94.1%	5.9%	100%
D, 0-4 yrs	Count	509	55	564
	%	90.2%	9.8%	100%
D, 5 yrs+	Count	616	54	670
	%	91.9%	8.1%	100%
S, 0-4 yrs	Count	383	34	417
	%	91.8%	8.2%	100%
Grade R; S, 5 yrs+; T	Count	805	36	841
	%	95.7%	4.3%	100%
Wave 8 Interviewer Age (5932 cases)				
less than 40 years	Count	166	29	195
	%	85.1%	14.9%	100%
40-49 years	Count	812	85	897

	%	90.5%	9.5%	100%
50-59 years	Count	2346	213	2559
	%	91.7%	8.3%	100%
60+ years	Count	2107	174	2281
	%	92.4%	7.6%	100%
Wave 8 Interviewer Age (7089 cases)				
less than 40 years	Count	265	37	302
	%	87.7%	12.3%	100%
40-49 years	Count	1607	125	1192
	%	89.5%	10.5%	100%
50-59 years	Count	2714	249	2963
	%	91.6%	8.43%	100%
60+ years	Count	2425	206	2631
	%	92.2%	7.8%	100%
Age of Youngest Child (5932 cases)				
No dependent children & 16-18 year olds	Count	551	63	614
	%	89.7%	10.3%	100%
0-4 year olds	Count	2312	214	2526
	%	91.5%	8.5%	100%
5-10 year olds	Count	1518	120	1638
	%	92.7%	7.3%	100%
11-15 year olds	Count	1050	104	1154
	%	91.0%	9.0%	100%
Age of Youngest Child (7089 cases)				
No dependent children & 16-18 year olds	Count	657	79	736
	%	89.3%	10.7%	100%
0-4 year olds	Count	2741	255	2996
	%	91.5%	8.5%	100%
5-10 year olds	Count	1793	150	1943
	%	92.3%	7.7%	100%
11-15 year olds	Count	1280	134	1414
	%	90.5%	9.5%	100%
Accommodation Type *(5932 cases)*				
Detached house	Count	1176	132	1308
	%	89.9%	10.1%	100%
Semi-detached house	Count	2204	188	2392
	%	92.1%	7.9%	100%
Terraced house	Count	1603	138	1741
	%	92.1%	7.9%	100%
Flat or maisonette - purpose built & Other	Count	397	38	435
	%	91.3%	8.7%	100%
Flat or maisonette - conversion	Count	51	5	56

	%	91.1%	8.9%	100%
Accommodation Type *(7089 cases)*				
Detached house	Count	1441	163	1604
	%	89.8%	10.2%	100%
Semi-detached house	Count	2575	227	2802
	%	91.9%	8.1%	100%
Terraced house	Count	1864	166	2030
	%	91.8%	8.2%	100%
Flat or maisonette – purpose built & Other	Count	509	56	565
	%	90.1%	9.9%	100%
Flat or maisonette – conversion	Count	82	6	88
	%	93.2%	6.8%	100%
Heating Problems in the Dwelling *(5932 cases)*				
Yes	Count	5073	474	5547
	%	91.5%	8.5%	100%
No & Don't know	Count	358	27	385
	%	93.0%	7.0%	100%
Heating Problems in the Dwelling *(7089 cases)*				
Yes	Count	6042	587	6629
	%	91.1%	8.9%	100%
No & Don't know	Count	429	31	460
	%	93.3%	6.7%	100%
Heating Problems in the Dwelling *(5932 cases)*				
Yes	Count	5073	474	5547
	%	91.5%	8.5%	100%
No & Don't know	Count	358	27	385
	%	93.0%	7.0%	100%
Heating Problems in the Dwelling *(7089 cases)*				
Yes	Count	6042	587	6629
	%	91.1%	8.9%	100%
No & Don't know	Count	429	31	460
	%	93.3%	6.7%	100%
Respondent Gender *(5932 cases)*				
Same interviewer	Count	3794	287	4081
	%	93.0%	7.0%	100%
More than one interviewer	Count	1637	214	1851
	%	88.4%	11.6%	100%
Respondent Gender *(7089 cases)*				
Same interviewer	Count	4491	351	4842
	%	92.8%	7.2%	100%
More than one interviewer	Count	1980	267	2247
	%	88.1%	11.9%	100%

Wave 8 Interviewer Gender (5932 cases)				
Male	Count	110	4	114
	%	96.5%	3.5%	100%
Female	Count	5321	497	5818
	%	91.5%	8.5%	100%
Wave 8 Interviewer Gender (7089 cases)				
Male	Count	125	7	132
	%	94.7%	5.3%	100%
Female	Count	6346	611	6957
	%	91.2%	8.8%	100%
Possession of any Academic or Vocational Qualification (5932 cases)				
No	Count	621	75	696
	%	89.2%	10.8%	100%
Yes	Count	4810	426	5236
	%	91.9%	8.1%	100%
Possession of any Academic or Vocational Qualification (7089 cases)				
No	Count	723	89	812
	%	89.0%	11.0%	100%
Yes	Count	5748	529	6277
	%	91.6%	8.4%	100%
Ethnicity (5932 cases)				
White	Count	5034	432	5466
	%	92.1%	7.9%	100%
Non-white and missing	Count	397	69	466
	%	85.2%	14.8%	100%
Ethnicity (7089 cases)				
White	Count	5987	534	6521
	%	91.8%	8.2%	100%
Non-white and missing	Count	484	84	568
	%	85.2%	14.8%	100%
First Wave for Respondent (5932 cases)				
Wave 1-Wave 4	Count	3955	315	4270
	%	92.6%	7.4%	100%
Wave 5-Wave 6	Count	903	90	993
	%	90.9%	9.1%	100%
Wave 7	Count	573	96	669
	%	85.7%	14.3%	100%
First Wave for Respondent (7089 cases)				
Wave 1-Wave 4	Count	4702	392	5094
	%	92.3%	7.7%	100%
Wave 5-Wave 6	Count	1070	109	1179
	%	90.8%	9.2%	100%

Wave 7	Count	699	117	816
	%	85.7%	14.3%	100%

VI.2. Appendix B – Data Generation for the Cross-classified Models

The data generation procedure defined below is specific to the medium scenario design, which 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 and the interviewer variance σ_u^2=0.3 and an overall probability π=0.8.

1. Create the interviewer effects in R and save them in an excel file.

```
#generate the interviewer random effects for B times and save them in an excel
file
#create a random normal variable 'u' of size k with mean 0 and standard
deviation equal to the square root of the variance sigmau2
#B is the number of simulations
#k is the number of interviewers
#sigmau2 is the interviewer-level variance
sim <- function(B=1000, k=240, sigmau2=0.3)
{
v1<-NULL
for(i in 1:B)
{
u <- rnorm(k,0,sqrt(sigmau2))
v1<-cbind(v1,u)
}
v1
}
data<-as.data.frame(sim())
write.csv(data, file="E:InterviewerEffects.csv")
write.table(data, file="E:InterviewerEffects.txt")
```

2. Delete the first row and first column from the excel file 'InterviewerEffects.csv' before using this file in the next step.

3. Create the area effects in R and save them in an excel file.

```
#generate the random area effects for B times and save them in an excel file
#B is the number of simulations
#I is the number of areas
#sigmau2 is the area-level variance
```

```
#create a random normal variable 'u' of size k with mean 0 and standard
deviation equal to the square root of the variance sigmav2
sim2 <- function(B=1000, l=120, sigmav2=0.3)
{
v2<-NULL
for(i in 1:B)
{
u <- rnorm(l,0,sqrt(sigmav2))
v2<-cbind(v2,u)
}
v2
}
data<-as.data.frame(sim2())
write.csv(data, file="E:AreaEffects.csv")
write.table(data, file="E:AreaEffects.txt")
```

4. Delete the first row and first column from the excel file ‘AreaEffects.csv’ before using this file in the next step.

5. Specify the interviewer identification code ‘INTid’ for each sample case for the CASE1 allocation scheme in an excel file and save it as "E: InterviewerALLOCATIONcase1.txt".

6. Simulate the dataset in R and save in an excel sheet.

```
#create a random variable of size n of 0s and 1s (as.numeric) which gives an
overall mean of pi
#n is the sample size
#pi is the overall probability of response
myrbin <- function(n, pi){as.numeric(runif(n) < pi)}
#simulate B samples from a cross-classified model with lp = beta0 + u + v and
replicate for B times the myrbin function
#work out the regression line lp by adding to the intercept beta0 the error
term for the interviewer v corresponding to the INTid and the error term for the
area u corresponding to the AREAid
#work out the probability of response ppi by calculating exp(lp)/1+exp(lp)
where lp comes from the previous line
#call myrbin of size k*m (number of interviewers multiplied by the number of
cases per interviewer) with probability equal to ppi to get a sample of 1s and
0s
#the area id AREAid is created by replicating n times the interviewer numbers
(1 to l)
#m is the number of cases per interviewer
#n is the number of cases per area
```

```
sim <- function(B=1000, k=240, l=120, m=24, n=48, beta0 = 1.39,
sigmau2=0.3, sigmav2=0.3)
 {
 y <- numeric(k*m)
 interviewerEFFECT<- numeric(l*n)
       areaEFFECT<-numeric(k*m)
        AREAid <- rep(1:l,rep(n,k))
        INTid <- as.list(read.table("E: InterviewerALLOCATIONcase1.txt", sep="",
header=FALSE))
        INTid <- INTid[[1]]
        u1<-as.data.frame(read.csv("E:AreaEffects.csv", sep=",", header=FALSE))
  v1<-as.data.frame(read.csv ("E:InterviewerEffects.csv", sep=",",
header=FALSE))
 for(i in 1:B){
 u <- u1[,i]
 v <- v1[,i]
 lp <- beta0 + rep(u[AREAid] + v[INTid])
 ppi <- exp(lp) /(1+exp(lp))
 y <- cbind(y, myrbin(n=k*m, pi=ppi))
 interviewerEFFECT<-cbind(interviewerEFFECT,v)
 areaEFFECT<-cbind(areaEFFECT,u)
 }
 cbind(y[,-1], AREAid, INTid, areaEFFECT[AREAid,-1], interviewerEFFECT[INTid,-
1])
 }
 data <- as.data.frame(sim())
 write.csv(data, file="E:\dataset1.csv")
```

7. Delete the first column from the excel file ‘dataset1.csv’ before using this file in the next step.

8. Open STATA. Click on File, Import, Text data created by a spreadsheet, Browse. Select ‘Comma Separated Values’ for file type and select ‘Dataset1.csv’. Click OK. Save as ‘dataset1.dta’.

9. Run the code below on the STATA datafile ‘datatset1.dta’. This sorts the data by the area identification and then by interviewer identification. This is important for running models in MLwiN. A serial number is created for each case. A variable cons, which is simply a string of 1s, is also created.

```
sort  areaid  intid
generate serialno=_n
generate cons=1
```

```
set matsize 11000
```

10. Using the same interviewer effects and area effects additional simulations are run for the other interviewer allocation schemes CASE2-CASE6.

VI.3. Appendix C - Model Estimation and Properties Calculations for the Cross-classified Models

The model estimation procedure defined below is specific to the medium scenario design, which 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 and the interviewer variance σ_u^2=0.3 and an overall probability π=0.8.

1. Open a STATA file on my computer and run the following code:

```
sysdir set PLUS S:\rv1g09\runmlwin
ssc install runmlwin
ssc install estout
adoupdate runmlwin
```

2. Go to Start, Programs, Accessories, Remote Desktop Connection. Write the following 'blue36.iridis.soton.ac.uk' and click Connect.

3. Once on 'blue36' which is the head node, remote desktop to a compute node.

4. Open the dataset 'dataset1.dta'.

5. Run the following code in the STATA dataset (make sure it is open with STATA12):

```
sysdir set PLUS S:\rv1g09\runmlwin
global MLwiN_path C:\Program Files (x86)\MLwiN v2.25\mlwin.exe
set matsize 11000
```

6. Fit the models in STATA by running the code below. Work is sent in batches of 100models (10 batches for every scenario). Save the results in an excel file on the S drive.

```
local i=1
while `i'<101 {
quietly runmlwin v`i' cons, level3 (intid:cons) level2 (areaid:cons) level1
(serialno:) discrete(distribution(binomial) link(logit) denominator(cons) pql2)
nopause maxiterations(150)
```

```
quietly runmlwin v`i' cons, level3 (intid:cons) level2 (areaid:cons) level1
(serialno:) discrete(distribution(binomial) link(logit) denominator(cons))
mcmc(cc burnin(10000) chain(200000)) initsprevious nopause
estimates store model`i'
local i= `i'+1
}
estout model1 model2 model3 model4 model5 model6 model7 model8
model9 model10 model11 model12 model13 model14 model15 model16
model17 model18 model19 model20 model21 model22 model23 model24
model25 model26 model27 model28 model29 model30 model31 model32
model33 model34 model35 model36 model37 model38 model39 model40
model41 model42 model43 model44 model45 model46 model47 model48
model49 model50 model51 model52 model53 model54 model55 model56
model57 model58 model59 model60 model61 model62 model63 model64
model65 model66 model67 model68 model69 model70 model71 model72
model73 model74 model75 model76 model77 model78 model79 model80
model81 model82 model83 model84 model85 model86 model87 model88
model89 model90 model91 model92 model93 model94 model95 model96
model97 model98 model99 model100, cells(b se ci_l ci_u ess meanmcse bd rl1
rl2 V[1] V[2] V[3] V[4] quantiles[2] quantiles[5] quantiles[8]) stats(N dic time
burnin chain converged), using "S:\output1-300dataset1.xls"
```

7. Delete irrelevant rows from each output excel sheet. Add the following variable names as the first column. Save file.

```
b0_b
b0_se
b0_min95
b0_max95
b0_ess
b0_meanmcse
b0_bd
b0_rl1
b0_rl2
b0_var11
b0_var12
b0_var13
b0_var14
b0_quantile2
b0_quantile5
b0_quantile8
area_b
area_se
```

area_min95
area_max95
area_ess
area_meanmcse
area_bd
area_rl1
area_rl2
area_var21
area_var22
area_var23
area_var24
area_quantile2
area_quantile5
area_quantile8
int_b
int_se
int_min95
int_max95
int_ess
int_meanmcse
int_bd
int_rl1
int_rl2
int_var31
int_var32
int_var33
int_var34
int_quantile2
int_quantile5
int_quantile8
N
Dic
Time
Burnin
Chain
Converged

8. Transpose rows with columns.

9. Merge the results from the 10 separate batches into one file with all 1000 models. Save excel file as '1000models_CASE1.xls'.

10. Open STATA. Click on File, Import, Excel Spreadsheet. Choose '1000models_CASE1.xls' from the Browse option. And click on Import First Row as Variable Names. Click Ok. Save the dataset as '1000models_ CASE1.dta'.

11. Run this code in the in the STATA file '1000models_ CASE1.dta' to obtain the various properties:

```
gen simulations=1000

#coverage rates based on the Wald test
gen waldCIcoverageAREAVAR=1
replace waldCIcoverageAREAVAR=0 if area_min95<0.3 & area_max95<0.3
replace waldCIcoverageAREAVAR=0 if area_min95>0.3 & area_max95>0.3
egen totalwaldCIcoverageAREAVAR=count(waldCIcoverageAREAVAR) if
waldCIcoverageAREAVAR==1

gen waldCIcoverageINTVAR=1
replace waldCIcoverageINTVAR=0 if int_min95<0.3 & int_max95<0.3
replace waldCIcoverageINTVAR=0 if int_min95>0.3 & int_max95>0.3
egen totalwaldCIcoverageINTVAR=count(waldCIcoverageINTVAR) if
waldCIcoverageINTVAR==1

gen waldCIcoverageINTERCEPT=1
replace waldCIcoverageINTERCEPT=0 if   b0_min95<1.39 &   b0_max95<1.39
replace waldCIcoverageINTERCEPT=0 if  b0_min95>1.39 &   b0_max95>1.39
egen totalwaldCIcoverageINTERCEPT=count(waldCIcoverageINTERCEPT) if
waldCIcoverageINTERCEPT==1
#coverage rates based on the MCMC credible intervals
gen mcmcCIcoverageAREAVAR=1
replace mcmcCIcoverageAREAVAR=0 if  area_quantile2<0.3 &
area_quantile8<0.3
replace mcmcCIcoverageAREAVAR=0 if  area_quantile2>0.3 &
area_quantile8>0.3
egen totalmcmcCIcoverageAREAVAR=count(mcmcCIcoverageAREAVAR) if
mcmcCIcoverageAREAVAR==1

gen mcmcCIcoverageINTVAR=1
replace mcmcCIcoverageINTVAR=0 if  int_quantile2<0.3 &  int_quantile8<0.3
replace mcmcCIcoverageINTVAR=0 if  int_quantile2>0.3 &  int_quantile8>0.3
egen totalmcmcCIcoverageINTVAR=count(mcmcCIcoverageINTVAR) if
mcmcCIcoverageINTVAR==1

gen mcmcCIcoverageINTERCEPT=1
replace mcmcCIcoverageINTERCEPT=0 if   b0_quantile2<1.39 &
b0_quantile8<1.39
replace mcmcCIcoverageINTERCEPT=0 if   b0_quantile2>1.39 &
b0_quantile8>1.39
egen totalmcmcCIcoverageINTERCEPT=count(mcmcCIcoverageINTERCEPT) if
mcmcCIcoverageINTERCEPT==1
```

```
#percentage relative biases based on mean and median
egen meanINTERCEPT=mean(b0_b)
gen biasINTERCEPT= (b0_b -1.39)/1.39*100
egen meanbiasINTERCEPT= mean(biasINTERCEPT)
egen meanINTERCEPTquantile5mcmc=mean(b0_quantile5)
gen biasINTERCEPTquantile5mcmc= (b0_quantile5-1.39)/1.39*100
egen meanbiasINTERCEPTquantile5mcmc =
mean(biasINTERCEPTquantile5mcmc)

egen meanAREAVAR=mean(area_b)
gen biasAREAVAR= (area_b -0.3)/0.3*100
egen meanbiasAREAVAR= mean(biasAREAVAR)
egen meanAREAVARquantile5mcmc=mean(area_quantile5)
gen biasAREAVARquantile5mcmc= (area_quantile5-0.3)/0.3*100
egen meanbiasAREAVARquantile5mcmc = mean(biasAREAVARquantile5mcmc)

egen meanINTVAR=mean(int_b)
gen biasINTVAR= (meanINTVAR-0.3)/0.3*100
egen meanbiasINTVAR= mean(biasINTVAR)
egen meanINTVARquantile5mcmc=mean(int_quantile5)
gen biasINTVARquantile5mcmc= (int_quantile5-0.3)/0.3*100
egen meanbiasINTVARquantile5mcmc = mean(biasINTVARquantile5mcmc)

#power of the Wald test at the 95% and 99% confidence levels
gen waldINTERCEPT=(b0_b/b0_se)^2
gen pvalINTERCEPT=chi2tail(1, waldINTERCEPT)
gen nullHacceptedINTERCEPT95=.
replace nullHacceptedINTERCEPT95=1 if pvalINTERCEPT>0.05
egen TOTALnullHacceptedINTERCEPT95= count(nullHacceptedINTERCEPT95)
gen powerINTERCEPT95=1-(TOTALnullHacceptedINTERCEPT95/ simulations)
gen nullHacceptedINTERCEPT99=.
replace nullHacceptedINTERCEPT99=1 if pvalINTERCEPT>0.01
egen TOTALnullHacceptedINTERCEPT99= count(nullHacceptedINTERCEPT99)
gen powerINTERCEPT99=1-(TOTALnullHacceptedINTERCEPT99/ simulations)

gen waldAREAVAR=(area_b/area_se)^2
gen pvalAREAVAR=[chi2tail(1, waldAREAVAR)]/2
gen nullHacceptedAREAVAR95=.
replace nullHacceptedAREAVAR95=1 if pvalAREAVAR>0.05
egen TOTALnullHacceptedAREAVAR95= count(nullHacceptedAREAVAR95)
gen powerAREAVAR95=1-(TOTALnullHacceptedAREAVAR95/ simulations)
gen nullHacceptedAREAVAR99=.
replace nullHacceptedAREAVAR99=1 if pvalAREAVAR>0.01
egen TOTALnullHacceptedAREAVAR99= count(nullHacceptedAREAVAR99)
gen powerAREAVAR99=1-(TOTALnullHacceptedAREAVAR99/simulations)

gen waldINTVAR=(int_b/int_se)^2
gen pvalINTVAR=[chi2tail(1, waldINTVAR)]/2
gen nullHacceptedINTVAR95=.
replace nullHacceptedINTVAR95=1 if pvalINTVAR>0.05
egen TOTALnullHacceptedINTVAR95= count(nullHacceptedINTVAR95)
```

```
gen powerINTVAR95=1-(TOTALnullHacceptedINTVAR95/ simulations)
gen nullHacceptedINTVAR99=.
replace nullHacceptedINTVAR99=1 if pvalINTVAR>0.01
egen TOTALnullHacceptedINTVAR99= count(nullHacceptedINTVAR99)
gen powerINTVAR99=1-(TOTALnullHacceptedINTVAR99/simulations)

# mean square error
gen mseINTERCEPT=(b0_b -1.39)^2
egen MEANmseINTERCEPT=mean(mseINTERCEPT)

gen mseAREAVAR=(area_b -0.3)^2
egen MEANmseAREAVAR=mean(mseAREAVAR)

gen mseINTVAR=(int_b -0.3)^2
egen MEANmseINTAVAR=mean(mseINTVAR)

#correlation of the two parameter estimators
egen MEANcovarAREAINT=mean(area_var23)
gen corrAREAINT= area_var23/(sqrt( area_var22)*sqrt( int_var33))
egen MEANcorrAREAINT=mean(corrAREAINT)

#% of times the values obtained for the Brooks-Draper and the Lower and
Upper Bound of the Raftery-Lewis diagnostics are less than the iteration length
specified
gen bdacceptedINTERCEPT=.
replace bdacceptedINTERCEPT =1 if b0_bd<200001
egen TOTALbdacceptedINTERCEPT = count(bdacceptedINTERCEPT)

gen bdacceptedAREAVAR=.
replace bdacceptedAREAVAR =1 if area_bd<200001
egen TOTALbdacceptedAREAVAR = count(bdacceptedAREAVAR)

gen bdacceptedINTVAR=.
replace bdacceptedINTVAR =1 if int_bd<200001
egen TOTALbdacceptedINTVAR = count(bdacceptedINTVAR)

gen rl1acceptedINTERCEPT=.
replace rl1acceptedINTERCEPT =1 if b0_rl1<200001
egen TOTALrl1acceptedINTERCEPT = count(rl1acceptedINTERCEPT)

gen rl1acceptedAREAVAR=.
replace rl1acceptedAREAVAR =1 if area_rl1<200001
egen TOTALrl1acceptedAREAVAR = count(rl1acceptedAREAVAR)

gen rl1acceptedINTVAR=.
replace rl1acceptedINTVAR =1 if int_rl1<200001
egen TOTALrl1acceptedINTVAR = count(rl1acceptedINTVAR)

gen rl2acceptedINTERCEPT=.
replace rl2acceptedINTERCEPT =1 if b0_rl2<200001
egen TOTALrl2acceptedINTERCEPT = count(rl2acceptedINTERCEPT)
```

```
gen rl2acceptedAREAVAR=.
replace rl2acceptedAREAVAR =1 if area_rl2<200001
egen TOTALrl2acceptedAREAVAR = count(rl2acceptedAREAVAR)

gen rl2acceptedINTVAR=.
replace rl2acceptedINTVAR =1 if int_rl2<200001
egen TOTALrl2acceptedINTVAR = count(rl2acceptedINTVAR)

#mean DIC
egen MEANdic=mean(dic)

#mean estimation running time
egen MEANtime=mean(time)

# monte carlo standard errors and respective confidence intervals
egen MEANmeanmcseINTERCEPT=mean(b0_meanmcse)
egen MEANmeanmcseAREA=mean(area_meanmcse)
egen MEANmeanmcseINT=mean(int_meanmcse)

gen MINmceINTERCEPT= b0_b - ( b0_se/sqrt(1000))
gen MAXmceINTERCEPT= b0_b + ( b0_se/sqrt(1000))
egen meanMAXmceINTERCEPT=mean(MAXmceINTERCEPT)
egen meanMINmceINTERCEPT=mean(MINmceINTERCEPT)
gen MINmceAREA= area_b - (area_se/sqrt(1000))
gen MAXmceAREA= area_b + (area_se/sqrt(1000))
egen meanMAXmceAREA=mean(MAXmceAREA)
egen meanMINmceAREA=mean(MINmceAREA)

gen MINmceINT= int_b - (int_se/sqrt(1000))
gen MAXmceINT= int_b + (int_se/sqrt(1000))
egen meanMAXmceINT=mean(MAXmceINT)
egen meanMINmceINT=mean(MINmceINT)

#standard errors
egen meanINTERCEPTse=mean(b0_se)
egen meanAREAVARse=mean(area_se)
egen meanINTVARse=mean(int_se)
```

VI.4. Appendix D – Relative Percentage Bias for Cross-Classified Models

The percentage relative bias rates based on the mean and the median values are presented below for the scenarios considered in the main paper. This data shows that generally the bias values obtained for the two measures are similar. One particular exception is the improved accuracy of estimators based on the median for scenarios with equal numbers of areas and interviewers and small sample sizes. These scenarios are in red text.

Percentage Relative Bias based on the Mean and Median Values for the Area and Interviewers Variance Estimators

CASE	Mean Area	Median Area	Mean Int	Median Int	N	N^A	N^I	σ_v^2	σ_u^2	Π
1	-3.2	-4.9	6.8	4.0	5760	120	240	0.3	0.3	0.8
2a	2.0	-0.1	1.3	-0.2	5760	120	240	0.3	0.3	0.8
2b	2.2	0.1	0.6	-0.8	5760	120	240	0.3	0.3	0.8
2c	2.4	0.3	-0.6	-2.0	5760	120	240	0.3	0.3	0.8
3a	2.4	0.3	0.1	-1.3	5760	120	240	0.3	0.3	0.8
3b	2.0	0.0	0.3	-1.1	5760	120	240	0.3	0.3	0.8
3c	1.2	-0.8	0.8	-0.6	5760	120	240	0.3	0.3	0.8
3d	2.2	0.2	1.0	-0.4	5760	120	240	0.3	0.3	0.8
3e	1.8	-0.2	0.1	-1.2	5760	120	240	0.3	0.3	0.8
3f	2.7	0.7	1.2	-0.1	5760	120	240	0.3	0.3	0.8
3h	1.9	-0.1	-0.7	-2.0	5760	120	240	0.3	0.3	0.8
4a	1.7	-0.2	0.7	-0.6	5760	120	240	0.3	0.3	0.8
4b	1.6	-0.4	0.6	-0.8	5760	120	240	0.3	0.3	0.8
4c	1.3	-0.6	0.6	-0.8	5760	120	240	0.3	0.3	0.8
5a	1.7	-0.2	1.0	-0.4	5760	120	240	0.3	0.3	0.8
5b	2.2	0.3	0.0	-1.3	5760	120	240	0.3	0.3	0.8
5c	2.3	0.4	-0.2	-1.5	5760	120	240	0.3	0.3	0.8
6a	1.1	-0.9	0.7	-0.6	5760	120	240	0.3	0.3	0.8
6b	1.8	-0.1	0.4	-0.9	5760	120	240	0.3	0.3	0.8
6c	1.7	-0.2	0.2	-1.1	5760	120	240	0.3	0.3	0.8
1	-6.7	-11.1	11.2	5.9	2880	60	120	0.3	0.3	0.8
2a	2.6	-1.8	1.9	-1.3	2880	60	120	0.3	0.3	0.8
2b	4.2	-0.2	1.3	-1.7	2880	60	120	0.3	0.3	0.8
2c	2.9	-1.3	2.4	-0.5	2880	60	120	0.3	0.3	0.8
3a	4.2	0.0	1.2	-1.6	2880	60	120	0.3	0.3	0.8

3b	2.6	-1.5	2.7	-0.1	2880	60	120	0.3	0.3	0.8
3c	3.5	-0.6	2.6	-0.2	2880	60	120	0.3	0.3	0.8
3d	3.6	-0.5	0.9	-1.9	2880	60	120	0.3	0.3	0.8
3e	3.3	-0.8	1.9	-0.9	2880	60	120	0.3	0.3	0.8
3f	3.9	-0.2	2.4	-0.4	2880	60	120	0.3	0.3	0.8
3h	3.0	-1.0	1.0	-1.7	2880	60	120	0.3	0.3	0.8
4a	3.3	-0.8	1.3	-1.5	2880	60	120	0.3	0.3	0.8
4b	3.8	-0.3	0.5	-2.2	2880	60	120	0.3	0.3	0.8
4c	3.8	-0.2	2.3	-0.5	2880	60	120	0.3	0.3	0.8
5a	2.4	-1.5	1.5	-1.2	2880	60	120	0.3	0.3	0.8
5b	2.6	-1.4	2.1	-0.7	2880	60	120	0.3	0.3	0.8
5c	3.6	-0.4	1.0	-1.7	2880	60	120	0.3	0.3	0.8
6a	3.1	-0.9	1.8	-0.9	2880	60	120	0.3	0.3	0.8
6b	3.0	-0.9	1.4	-1.3	2880	60	120	0.3	0.3	0.8
6c	4.3	0.3	0.9	-1.7	2880	60	120	0.3	0.3	0.8
E	2.8	-0.9	1.2	-1.3	2880	60	120	0.3	0.3	0.8
1	-5.3	-17.2	19.8	9.8	1440	30	60	0.3	0.3	0.8
2a	4.8	-4.4	2.4	-4.0	1440	30	60	0.3	0.3	0.8
2b	6.8	-2.3	4.5	-1.8	1440	30	60	0.3	0.3	0.8
2c	7.3	-1.9	5.6	-0.7	1440	30	60	0.3	0.3	0.8
3a	6.1	-2.0	1.1	-4.3	1440	30	60	0.3	0.3	0.8
3b	7.8	-1.0	4.4	-1.5	1440	30	60	0.3	0.3	0.8
3c	5.3	-3.2	0.4	-5.4	1440	30	60	0.3	0.3	0.8
3d	7.3	-1.3	2.0	-3.7	1440	30	60	0.3	0.3	0.8
3e	6.0	-2.5	0.1	-5.6	1440	30	60	0.3	0.3	0.8
3f	6.5	-2.0	1.4	-4.3	1440	30	60	0.3	0.3	0.8
3h	7.5	-1.1	2.2	-3.5	1440	30	60	0.3	0.3	0.8
4a	5.0	-3.5	1.8	-4.0	1440	30	60	0.3	0.3	0.8
4b	6.3	-2.1	0.8	-4.8	1440	30	60	0.3	0.3	0.8
4c	6.0	-2.4	2.1	-3.5	1440	30	60	0.3	0.3	0.8
5a	7.2	-1.2	3.4	-2.2	1440	30	60	0.3	0.3	0.8
5b	8.2	-0.2	1.5	-4.0	1440	30	60	0.3	0.3	0.8
5c	8.0	-0.3	1.0	-4.5	1440	30	60	0.3	0.3	0.8
6a	7.4	-0.9	2.4	-3.1	1440	30	60	0.3	0.3	0.8
6b	6.0	-2.1	1.0	-4.4	1440	30	60	0.3	0.3	0.8
6c	6.1	-2.0	1.1	-4.3	1440	30	60	0.3	0.3	0.8
E	5.9	-2.0	1.2	-4.0	1440	30	60	0.3	0.3	0.8
1	-4.4	-6.6	6.9	3.2	5760	120	240	0.3	0.3	0.9
2a	1.7	-0.9	-0.4	-2.4	5760	120	240	0.3	0.3	0.9
2c	0.4	-2.2	0.3	-1.7	5760	120	240	0.3	0.3	0.9
3a	1.6	-1.0	-0.5	-2.4	5760	120	240	0.3	0.3	0.9

3e	0.9	-1.6	-0.4	-2.2	5760	120	240	0.3	0.3	0.9
3h	0.7	-1.8	0.2	-1.7	5760	120	240	0.3	0.3	0.9
4a	1.3	-1.1	-0.5	-2.3	5760	120	240	0.3	0.3	0.9
4c	1.0	-1.4	-0.9	-2.7	5760	120	240	0.3	0.3	0.9
5a	1.8	-0.6	0.7	-1.2	5760	120	240	0.3	0.3	0.9
5c	2.1	-0.3	0.4	-1.4	5760	120	240	0.3	0.3	0.9
6a	1.2	-1.2	-0.5	-2.3	5760	120	240	0.3	0.3	0.9
6c	0.7	-1.7	0.7	-1.1	5760	120	240	0.3	0.3	0.9
1	-3.3	-4.8	5.4	2.9	5760	120	240	0.3	0.3	0.7
2a	2.1	0.2	1.0	-0.3	5760	120	240	0.3	0.3	0.7
2c	1.3	-0.6	0.9	-0.4	5760	120	240	0.3	0.3	0.7
3a	1.4	-0.5	0.7	-0.5	5760	120	240	0.3	0.3	0.7
3e	1.9	0.1	0.1	-1.1	5760	120	240	0.3	0.3	0.7
3h	1.9	0.1	0.6	-0.6	5760	120	240	0.3	0.3	0.7
4a	1.5	-0.3	0.9	-0.3	5760	120	240	0.3	0.3	0.7
4c	3.0	1.2	0.6	-0.5	5760	120	240	0.3	0.3	0.7
5a	2.1	0.3	1.5	0.3	5760	120	240	0.3	0.3	0.7
5c	2.3	0.6	0.7	-0.5	5760	120	240	0.3	0.3	0.7
6a	0.5	-1.2	1.0	-0.2	5760	120	240	0.3	0.3	0.7
6c	1.3	-0.4	1.5	0.3	5760	120	240	0.3	0.3	0.7
1	-5.9	-7.7	9.5	6.3	5760	120	240	0.2	0.2	0.8
2a	1.8	-0.5	0.6	-1.2	5760	120	240	0.2	0.2	0.8
2c	2.3	0.0	-0.1	-1.8	5760	120	240	0.2	0.2	0.8
3a	2.6	0.4	1.2	-0.5	5760	120	240	0.2	0.2	0.8
3e	1.6	-0.6	-1.4	-3.1	5760	120	240	0.2	0.2	0.8
3h	2.4	0.2	1.2	-0.4	5760	120	240	0.2	0.2	0.8
4a	1.6	-0.6	0.3	-1.3	5760	120	240	0.2	0.2	0.8
4c	1.3	-0.9	-0.1	-1.7	5760	120	240	0.2	0.2	0.8
5a	0.9	-1.3	0.0	-1.6	5760	120	240	0.2	0.2	0.8
5c	1.1	-1.1	-0.9	-2.5	5760	120	240	0.2	0.2	0.8
6a	1.9	-0.3	-0.1	-1.6	5760	120	240	0.2	0.2	0.8
6c	1.8	-0.4	-0.3	-1.8	5760	120	240	0.2	0.2	0.8
1	-2.0	-3.5	5.6	3.0	5760	120	240	0.4	0.4	0.8
2a	3.3	1.3	0.4	-1.0	5760	120	240	0.4	0.4	0.8
2c	1.4	-0.5	0.9	-0.4	5760	120	240	0.4	0.4	0.8
3a	1.9	0.0	0.9	-0.3	5760	120	240	0.4	0.4	0.8
3e	2.7	0.8	0.7	-0.5	5760	120	240	0.4	0.4	0.8
3h	2.1	0.2	0.6	-0.6	5760	120	240	0.4	0.4	0.8
4a	2.2	0.3	2.0	0.8	5760	120	240	0.4	0.4	0.8
4c	2.0	0.2	0.7	-0.5	5760	120	240	0.4	0.4	0.8

5a	2.7	0.9	2.0	0.8	5760	120	240	0.4	0.4	0.8
5c	2.6	0.8	1.3	0.1	5760	120	240	0.4	0.4	0.8
6a	2.4	0.6	0.4	-0.8	5760	120	240	0.4	0.4	0.8
6c	2.7	0.9	1.6	0.4	5760	120	240	0.4	0.4	0.8
1	-4.7	-6.5	7.5	5.1	5760	120	240	0.3	0.4	0.8
2a	2.2	0.0	1.2	-0.1	5760	120	240	0.3	0.4	0.8
2c	2.0	-0.2	0.6	-0.7	5760	120	240	0.3	0.4	0.8
3a	1.7	-0.4	0.1	-1.2	5760	120	240	0.3	0.4	0.8
3e	2.0	0.0	0.9	-0.4	5760	120	240	0.3	0.4	0.8
3h	1.1	-0.9	0.7	-0.5	5760	120	240	0.3	0.4	0.8
4a	3.0	1.0	0.7	-0.6	5760	120	240	0.3	0.4	0.8
4c	1.1	-0.9	0.7	-0.5	5760	120	240	0.3	0.4	0.8
5a	1.8	-0.2	0.3	-0.9	5760	120	240	0.3	0.4	0.8
5c	2.2	0.2	1.1	-0.1	5760	120	240	0.3	0.4	0.8
6a	1.7	-0.3	-0.3	-1.5	5760	120	240	0.3	0.4	0.8
6c	2.2	0.2	0.8	-0.4	5760	120	240	0.3	0.4	0.8
1	-0.4	-2.0	2.7	0.0	5760	120	240	0.4	0.3	0.8
2a	2.7	0.8	0.9	-0.6	5760	120	240	0.4	0.3	0.8
2c	2.5	0.6	0.6	-0.9	5760	120	240	0.4	0.3	0.8
3a	1.8	-0.1	1.5	0.1	5760	120	240	0.4	0.3	0.8
3e	2.8	0.9	-0.6	-2.0	5760	120	240	0.4	0.3	0.8
3h	2.1	0.3	-0.2	-1.6	5760	120	240	0.4	0.3	0.8
4a	2.3	0.5	0.5	-0.9	5760	120	240	0.4	0.3	0.8
4c	1.6	-0.2	1.9	0.5	5760	120	240	0.4	0.3	0.8
5a	2.6	0.8	0.4	-0.9	5760	120	240	0.4	0.3	0.8
5c	2.5	0.7	0.5	-0.8	5760	120	240	0.4	0.3	0.8
6a	1.9	0.1	0.7	-0.6	5760	120	240	0.4	0.3	0.8
6c	2.6	0.8	0.2	-1.1	5760	120	240	0.4	0.3	0.8
1	0.0	-1.7	2.7	-0.2	5760	120	240	0.3	0.2	0.8
2a	2.3	0.3	-0.2	-2.1	5760	120	240	0.3	0.2	0.8
2c	2.3	0.3	-1.3	-3.1	5760	120	240	0.3	0.2	0.8
3a	1.8	-0.2	-1.8	-3.4	5760	120	240	0.3	0.2	0.8
3e	2.1	0.1	0.4	-1.3	5760	120	240	0.3	0.2	0.8
3h	3.0	1.1	-1.8	-3.5	5760	120	240	0.3	0.2	0.8
4a	1.5	-0.4	-0.2	-1.9	5760	120	240	0.3	0.2	0.8
4c	2.2	0.3	0.7	-0.9	5760	120	240	0.3	0.2	0.8
5a	2.9	1.0	-0.7	-2.3	5760	120	240	0.3	0.2	0.8
5c	2.0	0.1	-0.7	-2.3	5760	120	240	0.3	0.2	0.8
6a	2.5	0.6	-0.7	-2.3	5760	120	240	0.3	0.2	0.8
6c	2.1	0.2	-1.2	-2.8	5760	120	240	0.3	0.2	0.8

1	-10.9	-13.4	7.8	5.3	5760	120	240	0.2	0.3	0.8
2a	-0.8	-3.3	1.4	0.0	5760	120	240	0.2	0.3	0.8
2c	0.1	-2.3	0.9	-0.5	5760	120	240	0.2	0.3	0.8
3a	0.3	-2.1	0.9	-0.5	5760	120	240	0.2	0.3	0.8
3e	-0.2	-2.5	1.0	-0.3	5760	120	240	0.2	0.3	0.8
3h	-0.5	-2.8	1.1	-0.2	5760	120	240	0.2	0.3	0.8
4a	0.6	-1.6	0.6	-0.7	5760	120	240	0.2	0.3	0.8
4c	0.2	-2.1	0.8	-0.5	5760	120	240	0.2	0.3	0.8
5a	0.2	-2.1	0.5	-0.9	5760	120	240	0.2	0.3	0.8
5c	0.2	-2.0	-0.2	-1.5	5760	120	240	0.2	0.3	0.8
6a	1.6	-0.6	-0.2	-1.5	5760	120	240	0.2	0.3	0.8
6c	0.7	-1.5	0.6	-0.7	5760	120	240	0.2	0.3	0.8
1	2.3	-4.7	3.6	-2.6	5760	120	120	0.3	0.3	0.8
2	3.6	1.1	1.5	-1.0	5760	120	120	0.3	0.3	0.8
3	1.6	-0.7	1.0	-1.4	5760	120	120	0.3	0.3	0.8
4	1.7	-0.3	1.9	-0.2	5760	120	120	0.3	0.3	0.8
5	2.0	0.0	1.4	-0.7	5760	120	120	0.3	0.3	0.8
6	1.6	-0.4	1.9	0.0	5760	120	120	0.3	0.3	0.8
1	4.4	-7.0	5.6	-4.4	2880	60	60	0.3	0.3	0.8
2	4.0	-0.9	5.0	0.0	2880	60	60	0.3	0.3	0.8
3	3.1	-1.6	4.3	-0.6	2880	60	60	0.3	0.3	0.8
4	1.5	-2.7	4.2	0.0	2880	60	60	0.3	0.3	0.8
5	2.6	-1.5	4.9	0.8	2880	60	60	0.3	0.3	0.8
6	3.8	-0.3	3.0	-1.0	2880	60	60	0.3	0.3	0.8
E	3.6	-0.2	3.9	0.2	2880	60	60	0.3	0.3	0.8
1	12.5	-6.5	11.3	-8.7	1440	30	30	0.3	0.3	0.8
2	10.8	0.1	9.0	-1.6	1440	30	30	0.3	0.3	0.8
3	10.5	0.4	5.3	-5.2	1440	30	30	0.3	0.3	0.8
4	9.8	0.6	9.7	0.5	1440	30	30	0.3	0.3	0.8
5	8.6	-0.2	8.3	-0.4	1440	30	30	0.3	0.3	0.8
6	10.3	1.6	6.7	-1.8	1440	30	30	0.3	0.3	0.8
E	9.1	1.0	8.2	0.2	1440	30	30	0.3	0.3	0.8

E represents the Extreme area/interviewer allocations

VI.5. Appendix E – Confidence Interval Coverage Rates for Cross-Classified Models

The confidence interval coverage rates based on the Wald and the coverage rates based on the MCMC credible intervals are presented below for the scenarios considered in the main paper. This data shows that there is lack of evidence that MCMC quantiles perform better than the Wald asymptotic normal in terms of the variance estimators coverage properties.

Wald and MCMC Credible Quantiles Confidence Interval Coverage Rates for the Area and Interviewers Variance Estimators

CASE	Wald Area	MCMC Area	Wald Int	MCMC Int	N	N^A	N^I	σ_v^2	σ_u^2	π
1	91.4	91.8	93.8	93.0	5760	120	240	0.3	0.3	0.8
2a	94.5	95.1	95.0	94.9	5760	120	240	0.3	0.3	0.8
2b	96.0	95.7	92.4	92.7	5760	120	240	0.3	0.3	0.8
2c	95.1	95.1	94.1	94.3	5760	120	240	0.3	0.3	0.8
3a	93.8	94.3	94.7	94.6	5760	120	240	0.3	0.3	0.8
3b	95.0	95.1	94.0	94.3	5760	120	240	0.3	0.3	0.8
3c	94.6	95.3	93.4	93.4	5760	120	240	0.3	0.3	0.8
3d	95.9	95.5	93.4	94.4	5760	120	240	0.3	0.3	0.8
3e	94.6	94.9	95.0	94.7	5760	120	240	0.3	0.3	0.8
3f	94.8	94.6	93.6	93.7	5760	120	240	0.3	0.3	0.8
3h	93.9	94.2	94.0	94.5	5760	120	240	0.3	0.3	0.8
4a	95.2	93.8	94.5	95.3	5760	120	240	0.3	0.3	0.8
4b	94.4	94.0	95.6	95.7	5760	120	240	0.3	0.3	0.8
4c	94.1	94.1	95.0	95.3	5760	120	240	0.3	0.3	0.8
5a	95.2	94.1	94.8	94.6	5760	120	240	0.3	0.3	0.8
5b	95.2	95.1	94.7	95.6	5760	120	240	0.3	0.3	0.8
5c	95.5	95.2	94.8	95.0	5760	120	240	0.3	0.3	0.8
6a	95.1	94.6	95.1	95.1	5760	120	240	0.3	0.3	0.8
6b	96.0	95.6	93.9	93.6	5760	120	240	0.3	0.3	0.8
6c	94.9	95.1	95.2	95.0	5760	120	240	0.3	0.3	0.8
1	90.1	91.8	93.6	93.0	2880	60	120	0.3	0.3	0.8
2a	92.9	93.2	93.5	93.9	2880	60	120	0.3	0.3	0.8
2b	94.0	93.9	94.1	94.9	2880	60	120	0.3	0.3	0.8
2c	93.3	93.7	92.8	93.9	2880	60	120	0.3	0.3	0.8
3a	92.8	92.3	94.3	94.5	2880	60	120	0.3	0.3	0.8
3b	94.1	94.1	94.3	94.4	2880	60	120	0.3	0.3	0.8

3c	94.1	94.0	93.8	94.1	2880	60	120	0.3	0.3	0.8
3d	93.0	94.0	94.1	94.5	2880	60	120	0.3	0.3	0.8
3e	93.7	94.1	94.4	94.7	2880	60	120	0.3	0.3	0.8
3f	95.0	93.8	95.6	94.8	2880	60	120	0.3	0.3	0.8
3h	94.1	94.5	93.1	93.7	2880	60	120	0.3	0.3	0.8
4a	94.5	94.3	93.0	94.3	2880	60	120	0.3	0.3	0.8
4b	94.0	94.7	93.5	93.7	2880	60	120	0.3	0.3	0.8
4c	94.5	93.8	95.5	95.8	2880	60	120	0.3	0.3	0.8
5a	94.8	94.7	93.6	94.1	2880	60	120	0.3	0.3	0.8
5b	94.3	93.3	93.8	94.4	2880	60	120	0.3	0.3	0.8
5c	94.1	93.6	94.9	94.7	2880	60	120	0.3	0.3	0.8
6a	93.6	93.3	94.6	94.6	2880	60	120	0.3	0.3	0.8
6b	93.9	94.1	94.0	94.9	2880	60	120	0.3	0.3	0.8
6c	94.9	94.4	94.5	94.4	2880	60	120	0.3	0.3	0.8
E	94.0	91.4	93.8	93.0	2880	60	120	0.3	0.3	0.8
1	87.7	91.6	91.0	91.0	1440	30	60	0.3	0.3	0.8
2a	91.2	92.0	91.1	92.9	1440	30	60	0.3	0.3	0.8
2b	92.9	93.4	91.8	92.6	1440	30	60	0.3	0.3	0.8
2c	92.6	93.6	91.0	92.2	1440	30	60	0.3	0.3	0.8
3a	93.7	94.6	92.5	94.3	1440	30	60	0.3	0.3	0.8
3b	92.7	93.4	92.7	92.8	1440	30	60	0.3	0.3	0.8
3c	92.8	93.0	89.9	90.7	1440	30	60	0.3	0.3	0.8
3d	92.7	93.9	91.5	93.0	1440	30	60	0.3	0.3	0.8
3e	92.4	93.7	91.1	93.4	1440	30	60	0.3	0.3	0.8
3f	93.3	93.6	92.0	92.1	1440	30	60	0.3	0.3	0.8
3h	93.4	93.4	91.2	92.6	1440	30	60	0.3	0.3	0.8
4a	92.9	93.3	91.2	92.6	1440	30	60	0.3	0.3	0.8
4b	92.3	92.1	91.3	93.9	1440	30	60	0.3	0.3	0.8
4c	92.7	92.5	92.6	94.3	1440	30	60	0.3	0.3	0.8
5a	94.1	94.7	92.7	93.8	1440	30	60	0.3	0.3	0.8
5b	93.1	91.7	93.6	94.0	1440	30	60	0.3	0.3	0.8
5c	92.9	93.1	91.8	92.8	1440	30	60	0.3	0.3	0.8
6a	93.5	93.4	91.5	92.8	1440	30	60	0.3	0.3	0.8
6b	93.5	94.1	92.0	93.3	1440	30	60	0.3	0.3	0.8
6c	93.7	94.6	92.5	94.3	1440	30	60	0.3	0.3	0.8
E	91.8	90.0	92.1	90.0	1440	30	60	0.3	0.3	0.8
1	91.5	92.0	93.7	94.3	5760	120	240	0.3	0.3	0.9
2a	94.2	94.4	93.1	93.9	5760	120	240	0.3	0.3	0.9
2c	94.9	94.4	93.0	93.9	5760	120	240	0.3	0.3	0.9
3a	94.3	94.3	93.2	93.8	5760	120	240	0.3	0.3	0.9
3e	94.6	94.6	94.3	94.9	5760	120	240	0.3	0.3	0.9

3h	94.9	94.5	95.0	94.9	5760	120	240	0.3	0.3	0.9
4a	94.1	94.0	93.7	93.6	5760	120	240	0.3	0.3	0.9
4c	95.0	95.0	94.7	94.2	5760	120	240	0.3	0.3	0.9
5a	94.7	95.7	93.9	94.2	5760	120	240	0.3	0.3	0.9
5c	94.9	95.7	94.5	94.4	5760	120	240	0.3	0.3	0.9
6a	94.2	94.4	92.8	93.2	5760	120	240	0.3	0.3	0.9
6c	93.7	93.6	95.0	94.8	5760	120	240	0.3	0.3	0.9
1	93.8	93.8	95.4	93.5	5760	120	240	0.3	0.3	0.7
2a	95.1	94.1	93.6	94.3	5760	120	240	0.3	0.3	0.7
2c	94.6	94.8	95.2	94.3	5760	120	240	0.3	0.3	0.7
3a	97.5	96.8	95.7	95.0	5760	120	240	0.3	0.3	0.7
3e	94.3	94.8	93.9	94.6	5760	120	240	0.3	0.3	0.7
3h	95.1	95.0	95.1	94.9	5760	120	240	0.3	0.3	0.7
4a	95.4	95.3	94.8	94.6	5760	120	240	0.3	0.3	0.7
4c	96.2	96.0	94.8	94.9	5760	120	240	0.3	0.3	0.7
5a	95.9	95.1	94.7	94.7	5760	120	240	0.3	0.3	0.7
5c	94.8	94.1	94.8	95.0	5760	120	240	0.3	0.3	0.7
6a	95.8	95.8	94.5	95.0	5760	120	240	0.3	0.3	0.7
6c	95.2	94.9	94.8	95.1	5760	120	240	0.3	0.3	0.7
1	91.3	92.4	94.3	93.5	5760	120	240	0.2	0.2	0.8
2a	94.3	94.7	93.8	94.1	5760	120	240	0.2	0.2	0.8
2c	94.2	93.2	94.6	94.3	5760	120	240	0.2	0.2	0.8
3a	94.2	94.8	93.6	93.2	5760	120	240	0.2	0.2	0.8
3e	94.1	94.6	93.2	93.4	5760	120	240	0.2	0.2	0.8
3h	94.1	94.4	95.6	95.5	5760	120	240	0.2	0.2	0.8
4a	95.1	94.3	94.4	94.4	5760	120	240	0.2	0.2	0.8
4c	93.5	93.7	93.8	94.4	5760	120	240	0.2	0.2	0.8
5a	94.7	95.0	94.0	94.8	5760	120	240	0.2	0.2	0.8
5c	93.6	93.9	94.7	95.0	5760	120	240	0.2	0.2	0.8
6a	93.6	93.3	93.7	94.2	5760	120	240	0.2	0.2	0.8
6c	93.7	93.7	95.2	95.4	5760	120	240	0.2	0.2	0.8
1	94.1	94.0	95.7	95.3	5760	120	240	0.4	0.4	0.8
2a	94.3	93.7	94.7	94.5	5760	120	240	0.4	0.4	0.8
2c	96.0	95.1	93.7	94.0	5760	120	240	0.4	0.4	0.8
3a	95.8	95.3	95.5	95.0	5760	120	240	0.4	0.4	0.8
3e	94.4	94.1	95.9	95.9	5760	120	240	0.4	0.4	0.8
3h	93.7	94.2	94.0	94.6	5760	120	240	0.4	0.4	0.8
4a	95.4	94.7	94.2	93.8	5760	120	240	0.4	0.4	0.8
4c	96.3	95.2	94.3	94.6	5760	120	240	0.4	0.4	0.8
5a	94.6	94.2	94.7	94.7	5760	120	240	0.4	0.4	0.8

5c	95.1	94.3	95.4	95.2	5760	120	240	0.4	0.4	0.8
6a	94.6	94.6	94.0	94.0	5760	120	240	0.4	0.4	0.8
6c	95.6	94.6	95.3	95.7	5760	120	240	0.4	0.4	0.8
1	91.6	93.3	93.9	92.1	5760	120	240	0.3	0.4	0.8
2a	95.0	95.0	95.0	94.7	5760	120	240	0.3	0.4	0.8
2c	93.3	94.1	93.6	93.8	5760	120	240	0.3	0.4	0.8
3a	94.4	94.6	93.7	94.1	5760	120	240	0.3	0.4	0.8
3e	94.8	95.1	95.2	94.9	5760	120	240	0.3	0.4	0.8
3h	94.2	95.4	94.9	95.5	5760	120	240	0.3	0.4	0.8
4a	95.8	94.6	94.3	94.4	5760	120	240	0.3	0.4	0.8
4c	94.9	95.0	95.5	94.7	5760	120	240	0.3	0.4	0.8
5a	94.7	94.3	94.9	93.9	5760	120	240	0.3	0.4	0.8
5c	94.1	95.0	94.1	93.5	5760	120	240	0.3	0.4	0.8
6a	94.3	94.3	94.2	95.1	5760	120	240	0.3	0.4	0.8
6c	95.9	95.8	94.2	94.7	5760	120	240	0.3	0.4	0.8
1	95.4	94.9	93.9	94.5	5760	120	240	0.4	0.3	0.8
2a	94.6	95.4	93.4	94.0	5760	120	240	0.4	0.3	0.8
2c	95.8	95.4	94.9	94.1	5760	120	240	0.4	0.3	0.8
3a	95.2	95.8	94.9	95.0	5760	120	240	0.4	0.3	0.8
3e	96.3	95.9	93.9	94.4	5760	120	240	0.4	0.3	0.8
3h	96.0	95.5	94.4	94.3	5760	120	240	0.4	0.3	0.8
4a	94.6	94.2	95.0	94.7	5760	120	240	0.4	0.3	0.8
4c	94.3	95.2	95.4	95.4	5760	120	240	0.4	0.3	0.8
5a	95.2	95.4	94.4	94.7	5760	120	240	0.4	0.3	0.8
5c	95.3	95.5	94.0	94.9	5760	120	240	0.4	0.3	0.8
6a	94.7	95.3	93.9	93.8	5760	120	240	0.4	0.3	0.8
6c	95.8	95.0	94.8	94.5	5760	120	240	0.4	0.3	0.8
1	95.2	93.9	93.5	94.1	5760	120	240	0.3	0.2	0.8
2a	94.8	95.0	94.3	94.7	5760	120	240	0.3	0.2	0.8
2c	95.4	94.6	93.7	94.0	5760	120	240	0.3	0.2	0.8
3a	95.4	95.5	94.0	95.0	5760	120	240	0.3	0.2	0.8
3e	95.4	95.2	93.3	93.9	5760	120	240	0.3	0.2	0.8
3h	93.7	93.2	92.2	93.9	5760	120	240	0.3	0.2	0.8
4a	94.8	94.5	93.2	93.3	5760	120	240	0.3	0.2	0.8
4c	94.2	93.6	94.7	94.5	5760	120	240	0.3	0.2	0.8
5a	94.4	94.6	94.3	95.2	5760	120	240	0.3	0.2	0.8
5c	96.0	95.5	93.2	93.6	5760	120	240	0.3	0.2	0.8
6a	94.6	94.4	94.6	95.6	5760	120	240	0.3	0.2	0.8
6c	95.7	95.2	94.2	94.4	5760	120	240	0.3	0.2	0.8

1	87.0	90.6	94.0	93.9	5760	120	240	0.2	0.3	0.8
2a	92.0	93.0	93.7	94.7	5760	120	240	0.2	0.3	0.8
2c	93.3	93.1	93.9	93.8	5760	120	240	0.2	0.3	0.8
3a	94.6	94.9	94.9	94.1	5760	120	240	0.2	0.3	0.8
3e	93.3	93.9	93.6	92.5	5760	120	240	0.2	0.3	0.8
3h	93.3	93.8	93.8	94.5	5760	120	240	0.2	0.3	0.8
4a	96.1	95.9	95.2	94.8	5760	120	240	0.2	0.3	0.8
4c	94.6	94.9	95.3	95.9	5760	120	240	0.2	0.3	0.8
5a	93.8	94.3	94.5	94.8	5760	120	240	0.2	0.3	0.8
5c	92.9	94.1	93.3	94.3	5760	120	240	0.2	0.3	0.8
6a	94.2	94.5	92.5	92.8	5760	120	240	0.2	0.3	0.8
6c	92.5	93.7	93.0	93.7	5760	120	240	0.2	0.3	0.8
1	99.7	99.7	99.7	99.7	5760	120	120	0.3	0.3	0.8
2	96.0	95.0	93.4	95.1	5760	120	120	0.3	0.3	0.8
3	94.7	94.8	93.7	94.5	5760	120	120	0.3	0.3	0.8
4	95.4	95.0	94.0	94.6	5760	120	120	0.3	0.3	0.8
5	95.4	96.2	94.2	94.8	5760	120	120	0.3	0.3	0.8
6	95.2	95.3	94.2	94.3	5760	120	120	0.3	0.3	0.8
1	100.0	100.0	99.9	100.0	2880	60	60	0.3	0.3	0.8
2	93.3	92.6	94.3	94.2	2880	60	60	0.3	0.3	0.8
3	94.9	95.4	94.4	93.3	2880	60	60	0.3	0.3	0.8
4	93.5	94.1	94.3	94.7	2880	60	60	0.3	0.3	0.8
5	94.2	94.9	95.2	94.7	2880	60	60	0.3	0.3	0.8
6	93.1	93.6	94.4	94.1	2880	60	60	0.3	0.3	0.8
E	94.1	94.5	94.2	94.9	2880	60	60	0.3	0.3	0.8
1	99.7	100.0	99.7	100.0	1440	30	30	0.3	0.3	0.8
2	92.0	92.8	91.3	93.1	1440	30	30	0.3	0.3	0.8
3	92.5	92.2	89.2	91.3	1440	30	30	0.3	0.3	0.8
4	94.1	94.8	93.1	93.3	1440	30	30	0.3	0.3	0.8
5	92.5	92.8	93.0	92.9	1440	30	30	0.3	0.3	0.8
6	93.6	93.9	91.8	92.8	1440	30	30	0.3	0.3	0.8
E	94.0	94.1	93.7	92.2	1440	30	30	0.3	0.3	0.8

E represents the Extreme area/interviewer allocations

VI.6. Appendix F – Data Generation for the Multiple Membership Models

The data generation procedure defined below is specific to the scenario with these factor specifications: N=5760, N_p^I=240, 24 cases per interviewer at the previous wave, σ_u^2=0.3, π=0.8 and a Type A change profile.

1. Create the interviewer effects in R and save them in an excel file.

```
#generate the interviewer random effects for B times and save them in an excel file
#create a random normal variable 'u' of size k with mean 0 and standard deviation equal to the square root of the variance sigmau2
#B is the number of simulations
#k is the number of interviewers
#sigmau2 is the interviewer-level variance
sim <- function(B=1000, k=240, sigmau2=0.3)
{
z1<-NULL
for(i in 1:B)
{
u <- rnorm(k,0,sqrt(sigmau2))
z1<-cbind(z1,u)
}
z1
}
data<-as.data.frame(sim())
write.csv(data, file="E: InterviewerEffects.csv")
write.table(data, file="E:InterviewerEffects.txt")
```

2. Delete the first row and first column from the excel file 'InterviewerEffects.csv' before using this file in the next step.

3. Generate the wave 1 interviewer allocations and save them in an excel sheet.

```
sim <- function(k=240, m=24)
{
x1<-NULL
clid <- rep(1:k,rep(m,k))
x1<-cbind(x1,clid)
}
data<-as.data.frame(sim())
write.csv(data, file="E:INTid1.csv")
write.table(data, file="E:INTid1.txt")
```

4. Delete the first row and first column from the excel file 'INTidW1.csv' before using this file in the next step.

5. Generate the probability of a case experiencing interviewer change between wave 1 and wave 2.

```
#Probability is the ratio of cases experiencing change to the ratio of cases
experiencing interviewer continuity for each interviewer
#m is the number of cases per interviewer at wave 1
#t is the number of cases to be selected from each wave 1 interviewer caseload
for re-allocation to a different interviewer at wave 2
Int.Change <- function(Probability=0.5, k=240, t=12, m=24)
{
  Vect.Prob        <- rep(Probability, times=m)
  Selected.Int.All  <- 9
  Selected.Int      <- rep(2, times=m)
  for (i in (1:k))
  {
  while (max(Selected.Int) >= 2)
    {
    Selected.Int   <- rmultinom(1, t, Vect.Prob)
    }
  Selected.Int.All <- c(Selected.Int.All,Selected.Int)
  Selected.Int     <- rep(2, times=m)
  }
  as.matrix(Selected.Int.All[-1], ncol=1)
}
IntChng <-Int.Change()
data<-as.data.frame(IntChng)
write.csv(data, file="E:IntChng.csv")
write.table(data, file="E:IntChng.txt")
```

6. Delete the first row and first column from the excel file 'IntChng.csv' before using this file in the next step.

7. Generate the wave 2 interviewer allocations and save them in an excel sheet.

```
sim2 <- function(k=240, t=12, m=24) {
x3<-NULL
clid <- rep(1:k,rep(m,k))
IntChng <- as.data.frame(read.csv("E:IntChng.csv", sep=",", header=FALSE))
clid3 <- rep(1:k, rep(t, k))
clid2 <- clid
clid2[IntChng==1] <- sample(clid3, t*k, replace=F)
```

```
x3<-cbind(x3,clid2)
}
data<-as.data.frame(sim2())
write.csv(data, file="E:INTid2.csv")
write.table(data, file="E:INTid2.txt")
```

8. Delete the first row and first column from the excel file 'INTid2.csv' before using this file in the next step.

9. Simulate the dataset in R and save in an excel sheet.

```
#create a random variable of size n of 0s and 1s (as.numeric) which gives an
overall mean of pi
#n is the sample size
#pi is the overall probability of response
myrbin <- function(n, pi){as.numeric(runif(n) < pi)}

# Simulate B samples from a multiple membership model with lp = beta0 +
w1*u_jc + w2*u_jp and replicate for B times the myrbin function
#work out the regression line lp by adding to the intercept beta0 the weighted
error terms of the previous and current wave interviewers corresponding to the
clid and clid2
#w1 and w2 are the model weights
#REALw1 and REALw2 are the real weights
        sim1 <- function(B=1000, k=240, m=24, beta0 = 1.39, sigmau2=0.3,
        w1=0.5, w2=0.5,  REALw1=0.5, REALw2=0.5)
                    {
                    y <- numeric(k*m)
                    interviewerW1EFFECT<- numeric(k*m)
                    interviewerW2EFFECT<- numeric(k*m)
                    clid <- as.list(read.csv("E:INTid1.csv", sep="", header=FALSE))
                    clid <- clid[[1]]
                    clid2 <- as.list(read.csv("E:INTid2.csv", sep="", header=FALSE))
              clid2 <- clid2 [[1]]
                    weight1 <- rep(w1, m*k)
                    weight2 <- rep(w2, m*k)
                    weight1[clid == clid2] <- 1
                    weight2[clid == clid2] <- 0
                    REALw1 <- rep(REALw1, m*k)
                    REALw2 <- rep(REALw2, m*k)
                    REALw1[clid== clid2] <- 1
                    REALw2[clid == clid2] <- 0
                    u1<-as.data.frame(read.csv("E: InterviewerEffects.csv",
        sep=",", header=FALSE))
                                for(i in 1:B)
```

```
                        {
                        u <- u1[,i]
                        lp <- beta0 + (REALw1 * u[clid]) + (REALw2 *
    u[clid2])
                        ppi <- exp(lp) /(1+exp(lp))
                        y <- cbind(y, myrbin(n=k*m, pi=ppi))
                        interviewer1EFFECT<-cbind(interviewer1EFFECT,u)
                        interviewer2EFFECT<-cbind(interviewer2EFFECT,u)
                        }
            cbind(y[,-1], clid, clid2, weight1, weight2, REALw1, REALw2,
    interviewer1EFFECT[clid,-1], interviewer2EFFECT[clid2,-1])
            }
    data <- as.data.frame(sim1())
    write.csv(data, file="E: dataset_realw1_0.5realw2_0.5-
    _modelw1_0.5_modelw2_0.5.csv")
```

10. Delete the first column from the excel file ‘dataset_realw1_0.5realw2_0.5_modelw1_0.5_modelw2_0.5.csv’ before using this file in the next step.

11. Open STATA. Click on File, Import, Text data created by a spreadsheet, Browse. Select ‘Comma Separated Values’ for file type and select ‘dataset_realw1_0.5realw2_0.5_modelw1_0.5_modelw2_0.5.csv’. Click OK. Save as ‘dataset_realw1_0.5realw2_0.5-_modelw1_0.5_modelw2_0.5.dta’.

12. Run the code below on the STATA datafile ‘datatset1.dta’. This changes the interviewer code for the wave 2 interviewer to 1 for cases with interviewer continuity, required for MLwiN to run MM models. This sorts the data by the two interviewer identification classifications. This is important for running models in MLwiN. A serial number is created for each case. A variable cons, which is simply a string of 1s, is also created.

```
replace clid2=0 if clid==clid2
sort  clid clid2
generate serialno=_n
generate cons=1
set matsize 11000
```

13. For the same dataset, apply different model weights (not the real weights). Save the new STATA file with the appropriate name (change model weights in title). To change the model weights the following code is run (for this example model weights are specified as 0.9 and 0.1)

```
replace weight1=0.9 if weight1==0.5
replace weight2=0.1 if weight2==0.5
```

VI.7. Appendix G – Procedure for Generating Interviewer Allocations under Different Change Profile Types for the Multiple Membership Models

The previous wave 1 interviewer allocations are constant across all profile types. The code specified in point 3 of Appendix F should be used for all profile type scenarios. The R code below generates the wave 2 interviewer allocations. The data generation procedure defined below is specific to the scenario with these factor specifications: N=5760, N_p^I=240, 24 cases per interviewer at the previous wave, σ_u^2=0.3, π=0.8.

- Type A

```
Int.Change <- function(Probability=0.5, Num.Ints=240,
Num.CasesToBeSelectedPerInt=12, Num.CasesPerInt=24)
{
  Vect.Prob         <- rep(Probability, times=Num.CasesPerInt)
  Selected.Int.All  <- 9
  Selected.Int      <- rep(2, times=Num.CasesPerInt)
  for (i in (1:Num.Ints))
  {
  while (max(Selected.Int) >= 2)
    {
    Selected.Int   <- rmultinom(1, Num.CasesToBeSelectedPerInt, Vect.Prob)
    }
  Selected.Int.All <- c(Selected.Int.All,Selected.Int)
  Selected.Int     <- rep(2, times=Num.CasesPerInt)
  }
  as.matrix(Selected.Int.All[-1], ncol=1)
}
IntChng <-Int.Change()
data<-as.data.frame(IntChng)
write.csv(data, file="E:IntChng.csv")
write.table(data, file="E:IntChng.txt")

sim2 <- function(Num.Ints=240, Num.CasesToBeSelectedPerInt=12,
Num.CasesToBeSelectedPerInt=24)
{
x3<-NULL
clid <- rep(1:Num.Ints,rep(Num.CasesToBeSelectedPerInt,Num.Ints))
```

```
IntChng <- as.data.frame(read.csv("E:IntChng.csv", sep=",", header=FALSE))
clid3 <- rep(1:Num.Ints, rep(Num.CasesToBeSelectedPerInt, Num.Ints))
clid2 <- clid
clid2[IntChng==1] <- sample(clid3, Num.CasesToBeSelectedPerInt*Num.Ints,
replace=F)
x3<-cbind(x3,clid2)
}
data<-as.data.frame(sim2())
write.csv(data, file="E:CLIDW2.csv")
write.table(data, file="E:CLIDW2.txt")
```

- Type B

```
Int.Change <- function(Probability=0.5, Num.Ints=240,
Num.CasesToBeSelectedPerInt=12, Num.CasesPerInt=24)
{
  Vect.Prob         <- rep(Probability, times=Num.CasesPerInt)
  Selected.Int.All  <- 9
  Selected.Int      <- rep(2, times=Num.CasesPerInt)
  for (i in (1:Num.Ints))
  {
  while (max(Selected.Int) >= 2)
    {
    Selected.Int   <- rmultinom(1, Num.CasesToBeSelectedPerInt, Vect.Prob)
    }
  Selected.Int.All <- c(Selected.Int.All,Selected.Int)
  Selected.Int     <- rep(2, times=Num.CasesPerInt)
  }
  as.matrix(Selected.Int.All[-1], ncol=1)
}
IntChng <-Int.Change()
data<-as.data.frame(IntChng)
write.csv(data, file="E:IntChng.csv")
write.table(data, file="E:IntChng.txt")

sim2 <- function(Num.Ints=240, Num.CasesToBeSelectedPerInt=12,
Num.CasesToBeSelectedPerInt=24)
{
x3<-NULL
clid <- rep(1:Num.Ints,rep(Num.CasesToBeSelectedPerInt,Num.Ints))
IntChng <- as.data.frame(read.csv("E:IntChng.csv", sep=",", header=FALSE))
clid3 <- rep(1:Num.Ints, rep(Num.CasesToBeSelectedPerInt, Num.Ints))
clid4 <-clid3 + 240
clid2 <- clid
```

```
clid2[IntChng==1] <- sample(clid4, Num.CasesToBeSelectedPerInt*Num.Ints,
replace=F)
x3<-cbind(x3,clid2)
}
data<-as.data.frame(sim2())
write.csv(data, file="E:CLIDW2.csv")
write.table(data, file="E:CLIDW2.txt")
```

- Type C

```
IntChng<-function(NoIntsSelected=Drops, Drops=120, Num.Ints=240,
Num.CasesPerInt=24, TotalN=5760)
{
  vector<-c(rep(0,Num.Ints))
  selected<-sample.int(Num.Ints,NoIntsSelected)
  vector[selected]<-1
  vectorFinal<-rep(0,Num.Ints*Num.CasesPerInt)
  vectorFinalId<-rep(0,Num.Ints*Num.CasesPerInt)
  for (i in 1:Num.Ints){
    if (vector[i]==1) {
      j<-(i-1)*Num.CasesPerInt+1
      vectorFinal[j:(j+Num.CasesPerInt-1)]<-1
         }
    Num.Ints<-(i-1)*Num.CasesPerInt+1
    vectorFinalId[Num.Ints:(Num.Ints+Num.CasesPerInt-1)]<-i
  }
  matrixFinal<-cbind(vectorFinalId,vectorFinal)
  matrixFinal
} data<-as.data.frame(IntChng())
write.csv(data, file="E:IntChng.csv")

sim2 <- function(Num.Ints=240, Num.CasesPerInt =24)
{
  clid <- rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints))
  f1<-IntChng()
  pool1 <- f1[which(f1[,2]==0)]
  pool2<-pool1[!duplicated(pool1)]
  pool <-rep(pool2, (Drops* Num.CasesPerInt)/(Num.Ints-Drops) )
  x2<-cbind(f1,clid)
  for (j in 1:(Num.Ints* Num.CasesPerInt)){
    if (x2[j,2]==1){
     if (length(pool)<2) {tmp<-pool
                          x2[j,3]<-tmp} else{
       tmp<-sample(pool,1)
```

```
            x2[j,3]<-tmp
            pool <- pool[-match(tmp, pool)] }
  }
  }
  x2[order(x2[,1],x2[,3]),]
}
data<-as.data.frame(sim2())
write.csv(data, file="E: CLIDW2.csv")
```

- Type D

```
IntChng<-function(NoIntsSelected=Drops, Drops=120, Num.Ints=240,
Num.CasesPerInt=24, TotalN=5760)
{
  vector<-c(rep(0,Num.Ints))
  selected<-sample.int(Num.Ints,NoIntsSelected)
  vector[selected]<-1
  vectorFinal<-rep(0,Num.Ints*Num.CasesPerInt)
  vectorFinalId<-rep(0,Num.Ints*Num.CasesPerInt)
  for (i in 1:Num.Ints){
    if (vector[i]==1) {
      j<-(i-1)*Num.CasesPerInt+1
      vectorFinal[j:(j+Num.CasesPerInt-1)]<-1
          }
    Num.Ints<-(i-1)*Num.CasesPerInt+1
    vectorFinalId[Num.Ints:(Num.Ints+Num.CasesPerInt-1)]<-i
  }
  matrixFinal<-cbind(vectorFinalId,vectorFinal)
  matrixFinal
}
data<-as.data.frame(IntChng())
write.csv(data, file="E:IntChng.csv")

sim2 <- function(Drops=120, Num.Ints=240, Num.CasesPerInt =24)
{
x3<-NULL
clid <- rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints))
IntChng <- as.data.frame(read.csv("E:IntChng.csv", sep=",", header=FALSE))
clid3 <- rep(1:Drops, rep(Num.CasesPerInt, Drops))
clid4 <-clid3 + 240
clid2 <- clid
clid2[IntChng==1] <- sample(clid4, Drops* Num.CasesPerInt, replace=F)
x3<-cbind(x3,clid2)
}
```

```
data<-as.data.frame(sim2())
write.csv(data, file="E:CLIDW2.csv")
```

- Type E

```
IntChng<-function(NoIntsSelected=Drops, Drops=120, Num.Ints=240,
Num.CasesPerInt=24, TotalN=5760)
{
  vector<-c(rep(0,Num.Ints))
  selected<-sample.int(Num.Ints,NoIntsSelected)
  vector[selected]<-1
  vectorFinal<-rep(0,Num.Ints*Num.CasesPerInt)
  vectorFinalId<-rep(0,Num.Ints*Num.CasesPerInt)
  for (i in 1:Num.Ints){
    if (vector[i]==1) {
      j<-(i-1)*Num.CasesPerInt+1
      vectorFinal[j:(j+Num.CasesPerInt-1)]<-1
         }
    Num.Ints<-(i-1)*Num.CasesPerInt+1
    vectorFinalId[Num.Ints:(Num.Ints+Num.CasesPerInt-1)]<-i
  }
  matrixFinal<-cbind(vectorFinalId,vectorFinal)
  matrixFinal
}
data<-as.data.frame(IntChng())
write.csv(data, file="E:IntChng.csv")

sim3 <- function(Num.Ints=240, Num.CasesPerInt =24)
{
  clid <- rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints))
  f1<-IntChng()
  pool1 <- f1[which(f1[,2]==0)]
  pool<-pool1[!duplicated(pool1)]
  x2<-unique(cbind(f1,clid))
  for (j in 1:Num.Ints){
    if (x2[j,2]==1){
     if (length(pool)<2) {tmp<-pool
                          x2[j,3]<-tmp} else{
       tmp<-sample(pool,1)
                 x2[j,3]<-tmp
                 pool <- pool[-match(tmp, pool)] }
  }
  }
```

```
   x2<-x2[order(x2[,1],x2[,3]),]
x2[rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints)),]
}
data<-as.data.frame(sim3())
write.csv(data, file="E: CLIDW2.csv")
```

- Type F

```
IntChng<-function(NoIntsSelected=Drops, Drops=120, Num.Ints=240,
Num.CasesPerInt=24, TotalN=5760)
{
  vector<-c(rep(0,Num.Ints))
  selected<-sample.int(Num.Ints,NoIntsSelected)
  vector[selected]<-1
  vectorFinal<-rep(0,Num.Ints*Num.CasesPerInt)
  vectorFinalId<-rep(0,Num.Ints*Num.CasesPerInt)
  for (i in 1:Num.Ints){
    if (vector[i]==1) {
      j<-(i-1)*Num.CasesPerInt+1
      vectorFinal[j:(j+Num.CasesPerInt-1)]<-1
          }
    Num.Ints<-(i-1)*Num.CasesPerInt+1
    vectorFinalId[Num.Ints:(Num.Ints+Num.CasesPerInt-1)]<-i
  }
  matrixFinal<-cbind(vectorFinalId,vectorFinal)
  matrixFinal
}
data<-as.data.frame(IntChng())
write.csv(data, file="E:IntChng.csv")

sim4 <- function(Num.Ints=240, Num.CasesPerInt =24)
{
  clid <- rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints))
  f1<-IntChng()
  pool1 <- f1[which(f1[,2]==0)]
  pool2<-pool1[!duplicated(pool1)]
  pool<-c((Num.Ints+1):(Num.Ints+length(pool2)))
x2<-unique(cbind(f1,clid))
  for (j in 1:Num.Ints){
    if (x2[j,2]==1){
     if (length(pool)<2) {tmp<-pool
                          x2[j,3]<-tmp} else{
       tmp<-sample(pool,1)
```

```
            x2[j,3]<-tmp
            pool <- pool[-match(tmp, pool)] }
  }
  }
    x2<-x2[order(x2[,1],x2[,3]),]
x2[rep(1:Num.Ints,rep(Num.CasesPerInt,Num.Ints)),]
}
data<-as.data.frame(sim4())
write.csv(data, file="E: CLIDW2.csv")
```

VI.8. Appendix H – Model Estimation and Properties Calculations for the Multiple Membership Models

The model estimation procedure defined below is specific to the scenario with these factor specifications: N=5760, N_p^I=240, 24 cases per interviewer at the previous wave, σ_u^2=0.3, π=0.8 and a Type A change profile.

1. Open a STATA file on my computer and run the following code:

```
sysdir set PLUS S:\rv1g09\runmlwin
ssc install runmlwin
ssc install estout
adoupdate runmlwin
```

2. Go to Start, Programs, Accessories, Remote Desktop Connection. Write the following 'blue36.iridis.soton.ac.uk' and click Connect.

3. Once on 'blue36' which is the head node, remote desktop to purple009 or purple010, which are the compute nodes.

4. Open the dataset 'dataset_realw1_0.5realw2_0.5-_modelw1_0.5_modelw2_0.5.dta' from an S drive file.

5. Run the following code in the STATA dataset (make sure it is open with STATA12)

```
sysdir set PLUS S:\rv1g09\runmlwin
global MLwiN_path C:\Program Files (x86)\MLwiN v2.25\mlwin.exe
set matsize 11000
```

6. Fit the models in STATA by running the code below. Work is sent in batches of 100models (10 batches for every scenario). Save the results in an excel file on the S drive.

```
local i=1
while `i'<101 {
quietly runmlwin v`i' cons, level2 (clid:cons) level1 (serialno:)
discrete(distribution(binomial) link(logit) denominator(cons) pql2) nopause
maxiterations(150)
quietly runmlwin v`i' cons, level2 (clid:cons, mmids(clid-clid2)
mmweights(weight1-weight2)) level1 (serialno:) discrete(distribution(binomial)
link(logit) denominator(cons)) mcmc(burnin(5000) chain(100000)) initsprevious
nopause
estimates store model`i'
```

```
local i=`i'+1
}
estout model1 model2 model3 model4 model5 model6 model7 model8
model9 model10 model11 model12 model13 model14 model15 model16
model17 model18 model19 model20 model21 model22 model23 model24
model25 model26 model27 model28 model29 model30 model31 model32
model33 model34 model35 model36 model37 model38 model39 model40
model41 model42 model43 model44 model45 model46 model47 model48
model49 model50 model51 model52 model53 model54 model55 model56
model57 model58 model59 model60 model61 model62 model63 model64
model65 model66 model67 model68 model69 model70 model71 model72
model73 model74 model75 model76 model77 model78 model79 model80
model81 model82 model83 model84 model85 model86 model87 model88
model89 model90 model91 model92 model93 model94 model95 model96
model97 model98 model99 model100, cells(b se ci_l ci_u ess meanmcse bd rl1
rl2 V[1] V[2] V[3] quantiles[2] quantiles[5] quantiles[8]) stats(N dic time burnin
chain converged), using " S:\ output1-300_realw1_0.5realw2_0.5-
_modelw1_0.5_modelw2_0.5.xls"
```

7. Delete irrelevant rows from each output excel sheet. Add the following variable names as the first column. Save file.

```
b0_b
b0_se
b0_min95
b0_max95
b0_ess
b0_meanmcse
b0_bd
b0_rl1
b0_rl2
b0_var11
b0_var12
b0_var13
b0_quantile2
b0_quantile5
b0_quantile8
int_b
int_se
int_min95
int_max95
int_ess
int_meanmcse
int_bd
int_rl1
```

int_rl2
int_var31
int_var32
int_var33
int_quantile2
int_quantile5
int_quantile8
N
Dic
Time
Burnin
Chain
Converged

8. Transpose rows with columns.

9. Merge the results from the 10 separate batches into one file with all 1000 models. Save excel file as '1000models_ realw1_0.5realw2_0.5-_modelw1_0.5_modelw2_0.5.xls'

10. Open STATA. Click on File, Import, Excel Spreadsheet. Choose '1000models_ realw1_0.5realw2_0.5¬_modelw1_0.5_modelw2_0.5.xls' from the Browse option. And click on Import First Row as Variable Names. Click Ok. Save the dataset as '1000models_ realw1_0.5realw2_0.5¬_modelw1_0.5_modelw2_0.5.dta'.

11. Add these columns to the dataset:

```
generate realweightW1=0.5
generate realweightW2=0.5
generate weightW1=0.5
generate weightW2=0.5
generate N_changes=2880
generate N_nochanges=2880
generate N_total=5760
generate ChangeRatio=12/24
generate ChangePercentage=12/24*100
generate NoChangeRatio=12/24
```

12. Run this code in the in the STATA file '1000models_ realw1_0.5realw2_0.5¬_modelw1_0.5_modelw2_0.5.dta' to obtain the various properties:

```
gen simulations=1000

#coverage rates based on the Wald test
```

```
gen waldCIcoverageINTVAR=1
replace waldCIcoverageINTVAR=0 if int_min95<0.3 & int_max95<0.3
replace waldCIcoverageINTVAR=0 if int_min95>0.3 & int_max95>0.3
egen totalwaldCIcoverageINTVAR=count(waldCIcoverageINTVAR) if
waldCIcoverageINTVAR==1

gen waldCIcoverageINTERCEPT=1
replace waldCIcoverageINTERCEPT=0 if   b0_min95<1.39 &   b0_max95<1.39
replace waldCIcoverageINTERCEPT=0 if  b0_min95>1.39 &   b0_max95>1.39
egen totalwaldCIcoverageINTERCEPT=count(waldCIcoverageINTERCEPT) if
waldCIcoverageINTERCEPT==1

#coverage rates based on the MCMC credible intervals
gen mcmcCIcoverageINTVAR=1
replace mcmcCIcoverageINTVAR=0 if  int_quantile2<0.3 &  int_quantile8<0.3
replace mcmcCIcoverageINTVAR=0 if  int_quantile2>0.3 &  int_quantile8>0.3
egen totalmcmcCIcoverageINTVAR=count(mcmcCIcoverageINTVAR) if
mcmcCIcoverageINTVAR==1

gen mcmcCIcoverageINTERCEPT=1
replace mcmcCIcoverageINTERCEPT=0 if   b0_quantile2<1.39 &
b0_quantile8<1.39
replace mcmcCIcoverageINTERCEPT=0 if   b0_quantile2>1.39 &
b0_quantile8>1.39
egen totalmcmcCIcoverageINTERCEPT=count(mcmcCIcoverageINTERCEPT) if
mcmcCIcoverageINTERCEPT==1

#percentage relative biases based on mean and median
egen meanINTERCEPT=mean(b0_b)
gen biasINTERCEPT= (b0_b -1.39)/1.39*100
egen meanbiasINTERCEPT= mean(biasINTERCEPT)
egen meanINTERCEPTquantile5mcmc=mean(b0_quantile5)
gen biasINTERCEPTquantile5mcmc= (b0_quantile5-1.39)/1.39*100
egen meanbiasINTERCEPTquantile5mcmc =
mean(biasINTERCEPTquantile5mcmc)

egen meanINTVAR=mean(int_b)
gen biasINTVAR= (meanINTVAR-0.3)/0.3*100
egen meanbiasINTVAR= mean(biasINTVAR)
egen meanINTVARquantile5mcmc=mean(int_quantile5)
gen biasINTVARquantile5mcmc= (int_quantile5-0.3)/0.3*100
egen meanbiasINTVARquantile5mcmc = mean(biasINTVARquantile5mcmc)

#power of the Wald test at the 95% and 99% confidence levels
```

```
gen waldINTERCEPT=(b0_b/b0_se)^2
gen pvalINTERCEPT=chi2tail(1, waldINTERCEPT)
gen nullHacceptedINTERCEPT95=.
replace nullHacceptedINTERCEPT95=1 if pvalINTERCEPT>0.05
egen TOTALnullHacceptedINTERCEPT95= count(nullHacceptedINTERCEPT95)
gen powerINTERCEPT95=1-(TOTALnullHacceptedINTERCEPT95/ simulations)

gen nullHacceptedINTERCEPT99=.
replace nullHacceptedINTERCEPT99=1 if pvalINTERCEPT>0.01
egen TOTALnullHacceptedINTERCEPT99= count(nullHacceptedINTERCEPT99)
gen powerINTERCEPT99=1-(TOTALnullHacceptedINTERCEPT99/ simulations)

gen waldINTVAR=(int_b/int_se)^2
gen pvalINTVAR=[chi2tail(1, waldINTVAR)]/2

gen nullHacceptedINTVAR95=.
replace nullHacceptedINTVAR95=1 if pvalINTVAR>0.05
egen TOTALnullHacceptedINTVAR95= count(nullHacceptedINTVAR95)
gen powerINTVAR95=1-(TOTALnullHacceptedINTVAR95/ simulations)

gen nullHacceptedINTVAR99=.
replace nullHacceptedINTVAR99=1 if pvalINTVAR>0.01
egen TOTALnullHacceptedINTVAR99= count(nullHacceptedINTVAR99)
gen powerINTVAR99=1-(TOTALnullHacceptedINTVAR99/simulations)

# mean square error
gen mseINTERCEPT=(b0_b -1.39)^2
egen MEANmseINTERCEPT=mean(mseINTERCEPT)

gen mseINTVAR=(int_b -0.3)^2
egen MEANmseINTAVAR=mean(mseINTVAR)
#% of times the values obtained for the Brooks-Draper and the Lower and
Upper Bound of the Raftery-Lewis diagnostics are less than the iteration length
specified
gen bdacceptedINTERCEPT=.
replace bdacceptedINTERCEPT =1 if b0_bd<100001
egen TOTALbdacceptedINTERCEPT = count(bdacceptedINTERCEPT)
gen bdacceptedINTVAR=.
replace bdacceptedINTVAR =1 if int_bd<100001
egen TOTALbdacceptedINTVAR = count(bdacceptedINTVAR)

gen rl1acceptedINTERCEPT=.
replace rl1acceptedINTERCEPT =1 if b0_rl1<100001
egen TOTALrl1acceptedINTERCEPT = count(rl1acceptedINTERCEPT)
```

```
gen rl1acceptedINTVAR=.
replace rl1acceptedINTVAR =1 if int_rl1<100001
egen TOTALrl1acceptedINTVAR = count(rl1acceptedINTVAR)

gen rl2acceptedINTERCEPT=.
replace rl2acceptedINTERCEPT =1 if b0_rl2<100001
egen TOTALrl2acceptedINTERCEPT = count(rl2acceptedINTERCEPT)

gen rl2acceptedINTVAR=.
replace rl2acceptedINTVAR =1 if int_rl2<100001
egen TOTALrl2acceptedINTVAR = count(rl2acceptedINTVAR)

#mean DIC
egen MEANdic=mean(dic)

#mean estimation running time
egen MEANtime=mean(time)

# monte carlo standard errors and respective confidence intervals
egen MEANmeanmcseINTERCEPT=mean(b0_meanmcse)
egen MEANmeanmcseINT=mean(int_meanmcse)

gen MINmceINTERCEPT= b0_b - ( b0_se/sqrt(1000))
gen MAXmceINTERCEPT= b0_b + ( b0_se/sqrt(1000))
egen meanMAXmceINTERCEPT=mean(MAXmceINTERCEPT)
egen meanMINmceINTERCEPT=mean(MINmceINTERCEPT)

gen MINmceINT= int_b - (int_se/sqrt(1000))
gen MAXmceINT= int_b + (int_se/sqrt(1000))
egen meanMAXmceINT=mean(MAXmceINT)
egen meanMINmceINT=mean(MINmceINT)

#standard errors
egen meanINTERCEPTse=mean(b0_se)
egen meanINTVARse=mean(int_se)
```

18. Once all models with the different weight combinations are fitted, the file with model weights set as 0.9 0.1 should be opened, and the other files (with different model weight specifications for the same simulated dataset) should be appended by clicking on Data, Combine datasets, Append datasets. Save dataset as '9000models_ realw1_0.5realw2_0.5_modelw1_0.5_modelw2_0.5.dta'.
19. Run the following STATA code to the dataset including all 9000 models:

```
# bestfit is a measure identifying the weighting scheme distribution for the
1000 models (out of 9000 possible models) obtaining the lowest DIC
egen simulation= seq(), f(1) t(1000)
sort weightW1
egen weightscheme= seq(), f(1) t(9) b(1000)
sort simulation
by simulation: egen minDIC=min(dic)
generate roundminDIC=round(minDIC, 0.001)
generate roundDIC=round(dic, 0.001)
generate bestfit=weightscheme if roundminDIC==roundDIC
tab bestfit if bestfit!=.

#calculate the percentage relative bias, power of the Wald test, standard errors
and confidence interval coverage for the 1000 models out of the total 9000
models which obtain the lowest DIC
egen meanINTVARbestfit=mean(int_b) if bestfit==weightscheme
gen biasINTVARbestfit=(int_b-0.3)/0.3*100 if bestfit==weightscheme
egen meanbiasINTVARbestfit= mean(biasINTVARbestfit) if
bestfit==weightscheme

egen simulationsBESTFIT=count(bestfit) if bestfit>0

gen waldCIcoverageINTVARbestfit=1 if bestfit==weightscheme
replace waldCIcoverageINTVARbestfit=0 if int_min95<0.3 & int_max95<0.3 &
bestfit==weightscheme
replace waldCIcoverageINTVARbestfit=0 if int_min95>0.3 & int_max95>0.3 &
bestfit==weightscheme
egen totalwaldCIcoverageINTVARbestfit=count(waldCIcoverageINTVARbestfit) if
waldCIcoverageINTVARbestfit==1 & bestfit==weightscheme
gen
totalwaldCIcovINTVARbestfitP=totalwaldCIcoverageINTVARbestfit/simulationsB
ESTFIT*100

gen nullHacceptedINTVAR95bestfit=. if bestfit==weightscheme
replace nullHacceptedINTVAR95bestfit=1 if pvalINTVAR>0.05 &
bestfit==weightscheme
egen TOTALnullHaccINTVAR95bestfit= count(nullHacceptedINTVAR95) if
bestfit==weightscheme
gen powerINTVAR95bestfit=1-
(TOTALnullHaccINTVAR95bestfit/simulationsBESTFIT) if
bestfit==weightscheme

egen meanINTVARsebestfit=mean(int_se) if bestfit==weightscheme
```

VI.9. Appendix I – Mean DIC Values for Multiple Membership Models

The mean DIC values for scenarios considered in the main paper are presented below. This measure is calculated for each model weights specification of each scenario, averaged over the 1000 simulations.

Mean DIC Values across Different Scenarios and Model Weights Specifications

Change Profile Type	DIC	N	% change	W_{ij}	w_{ij}
A	2910.7	2880	8	0.5, 0.5	0.9, 0.1
A	2909.5	2880	8	0.5, 0.5	0.8, 0.2
A	2908.7	2880	8	0.5, 0.5	0.7, 0.3
A	2908.2	2880	8	0.5, 0.5	0.6, 0.4
A	2908.0	2880	8	0.5, 0.5	0.5, 0.5
A	2908.3	2880	8	0.5, 0.5	0.4, 0.6
A	2908.8	2880	8	0.5, 0.5	0.3, 0.7
A	2909.8	2880	8	0.5, 0.5	0.2, 0.8
A	2911.0	2880	8	0.5, 0.5	0.1, 0.9
A	2908.9	2880	8	0.9, 0.1	0.9, 0.1
A	2909.0	2880	8	0.9, 0.1	0.8, 0.2
A	2909.5	2880	8	0.9, 0.1	0.7, 0.3
A	2910.4	2880	8	0.9, 0.1	0.6, 0.4
A	2911.7	2880	8	0.9, 0.1	0.5, 0.5
A	2913.3	2880	8	0.9, 0.1	0.4, 0.6
A	2915.2	2880	8	0.9, 0.1	0.3, 0.7
A	2917.4	2880	8	0.9, 0.1	0.2, 0.8
A	2919.8	2880	8	0.9, 0.1	0.1, 0.9
A	2924.7	2880	50	0.5, 0.5	0.9, 0.1
A	2920.0	2880	50	0.5, 0.5	0.8, 0.2
A	2916.1	2880	50	0.5, 0.5	0.7, 0.3
A	2913.4	2880	50	0.5, 0.5	0.6, 0.4
A	2912.5	2880	50	0.5, 0.5	0.5, 0.5
A	2913.5	2880	50	0.5, 0.5	0.4, 0.6
A	2916.2	2880	50	0.5, 0.5	0.3, 0.7

A	2920.3	2880	50	0.5, 0.5	0.2, 0.8
A	2925.0	2880	50	0.5, 0.5	0.1, 0.9
A	2914.4	2880	50	0.9, 0.1	0.9, 0.1
A	2915.3	2880	50	0.9, 0.1	0.8, 0.2
A	2918.3	2880	50	0.9, 0.1	0.7, 0.3
A	2923.5	2880	50	0.9, 0.1	0.6, 0.4
A	2930.5	2880	50	0.9, 0.1	0.5, 0.5
A	2938.8	2880	50	0.9, 0.1	0.4, 0.6
A	2947.2	2880	50	0.9, 0.1	0.3, 0.7
A	2955.1	2880	50	0.9, 0.1	0.2, 0.8
A	2961.5	2880	50	0.9, 0.1	0.1, 0.9
A	5825.4	5760	8	0.5, 0.5	0.9, 0.1
A	5823.0	5760	8	0.5, 0.5	0.8, 0.2
A	5821.3	5760	8	0.5, 0.5	0.7, 0.3
A	5820.2	5760	8	0.5, 0.5	0.6, 0.4
A	5819.9	5760	8	0.5, 0.5	0.5, 0.5
A	5820.2	5760	8	0.5, 0.5	0.4, 0.6
A	5821.3	5760	8	0.5, 0.5	0.3, 0.7
A	5823.1	5760	8	0.5, 0.5	0.2, 0.8
A	5825.5	5760	8	0.5, 0.5	0.1, 0.9
A	5826.2	5760	8	0.7, 0.3	0.9, 0.1
A	5825.1	5760	8	0.7, 0.3	0.8, 0.2
A	5824.7	5760	8	0.7, 0.3	0.7, 0.3
A	5825.1	5760	8	0.7, 0.3	0.6, 0.4
A	5826.1	5760	8	0.7, 0.3	0.5, 0.5
A	5827.9	5760	8	0.7, 0.3	0.4, 0.6
A	5830.3	5760	8	0.7, 0.3	0.3, 0.7
A	5833.4	5760	8	0.7, 0.3	0.2, 0.8
A	5837.0	5760	8	0.7, 0.3	0.1, 0.9
A	5826.1	5760	8	0.9, 0.1	0.9, 0.1
A	5826.4	5760	8	0.9, 0.1	0.8, 0.2
A	5827.5	5760	8	0.9, 0.1	0.7, 0.3
A	5829.3	5760	8	0.9, 0.1	0.6, 0.4
A	5831.8	5760	8	0.9, 0.1	0.5, 0.5
A	5835.0	5760	8	0.9, 0.1	0.4, 0.6
A	5838.7	5760	8	0.9, 0.1	0.3, 0.7
A	5843.1	5760	8	0.9, 0.1	0.2, 0.8
A	5847.9	5760	8	0.9, 0.1	0.1, 0.9

A	5837.0	5760	21	0.5, 0.5	0.9, 0.1
A	5831.6	5760	21	0.5, 0.5	0.8, 0.2
A	5827.5	5760	21	0.5, 0.5	0.7, 0.3
A	5824.9	5760	21	0.5, 0.5	0.6, 0.4
A	5824.1	5760	21	0.5, 0.5	0.5, 0.5
A	5824.9	5760	21	0.5, 0.5	0.4, 0.6
A	5827.5	5760	21	0.5, 0.5	0.3, 0.7
A	5831.6	5760	21	0.5, 0.5	0.2, 0.8
A	5837.0	5760	21	0.5, 0.5	0.1, 0.9
A	5832.1	5760	33	0.5, 0.5	0.9, 0.1
A	5824.6	5760	33	0.5, 0.5	0.8, 0.2
A	5818.6	5760	33	0.5, 0.5	0.7, 0.3
A	5814.8	5760	33	0.5, 0.5	0.6, 0.4
A	5813.6	5760	33	0.5, 0.5	0.5, 0.5
A	5815.1	5760	33	0.5, 0.5	0.4, 0.6
A	5819.1	5760	33	0.5, 0.5	0.3, 0.7
A	5825.3	5760	33	0.5, 0.5	0.2, 0.8
A	5833.0	5760	33	0.5, 0.5	0.1, 0.9
A	5854.2	5760	50	0.5, 0.5	0.9, 0.1
A	5844.7	5760	50	0.5, 0.5	0.8, 0.2
A	5836.6	5760	50	0.5, 0.5	0.7, 0.3
A	5831.0	5760	50	0.5, 0.5	0.6, 0.4
A	5829.1	5760	50	0.5, 0.5	0.5, 0.5
A	5831.0	5760	50	0.5, 0.5	0.4, 0.6
A	5836.5	5760	50	0.5, 0.5	0.3, 0.7
A	5844.6	5760	50	0.5, 0.5	0.2, 0.8
A	5854.2	5760	50	0.5, 0.5	0.1, 0.9
A	5826.9	5760	50	0.7, 0.3	0.9, 0.1
A	5822.0	5760	50	0.7, 0.3	0.8, 0.2
A	5820.1	5760	50	0.7, 0.3	0.7, 0.3
A	5821.9	5760	50	0.7, 0.3	0.6, 0.4
A	5827.7	5760	50	0.7, 0.3	0.5, 0.5
A	5837.1	5760	50	0.7, 0.3	0.4, 0.6
A	5848.8	5760	50	0.7, 0.3	0.3, 0.7
A	5861.5	5760	50	0.7, 0.3	0.2, 0.8
A	5873.9	5760	50	0.7, 0.3	0.1, 0.9
A	5816.9	5760	50	0.9, 0.1	0.9, 0.1
A	5818.7	5760	50	0.9, 0.1	0.8, 0.2
A	5825.0	5760	50	0.9, 0.1	0.7, 0.3

A	5835.2	5760	50	0.9, 0.1	0.6, 0.4
A	5849.5	5760	50	0.9, 0.1	0.5, 0.5
A	5866.1	5760	50	0.9, 0.1	0.4, 0.6
A	5883.2	5760	50	0.9, 0.1	0.3, 0.7
A	5899.1	5760	50	0.9, 0.1	0.2, 0.8
A	5912.7	5760	50	0.9, 0.1	0.1, 0.9
A	5850.1	5760	92	0.5, 0.5	0.9, 0.1
A	5840.9	5760	92	0.5, 0.5	0.8, 0.2
A	5831.3	5760	92	0.5, 0.5	0.7, 0.3
A	5823.6	5760	92	0.5, 0.5	0.6, 0.4
A	5820.7	5760	92	0.5, 0.5	0.5, 0.5
A	5823.9	5760	92	0.5, 0.5	0.4, 0.6
A	5831.8	5760	92	0.5, 0.5	0.3, 0.7
A	5841.6	5760	92	0.5, 0.5	0.2, 0.8
A	5850.7	5760	92	0.5, 0.5	0.1, 0.9
B	2918.7	2880	8	0.5, 0.5	0.9, 0.1
B	2918.1	2880	8	0.5, 0.5	0.8, 0.2
B	2917.7	2880	8	0.5, 0.5	0.7, 0.3
B	2917.4	2880	8	0.5, 0.5	0.6, 0.4
B	2917.3	2880	8	0.5, 0.5	0.5, 0.5
B	2917.4	2880	8	0.5, 0.5	0.4, 0.6
B	2917.6	2880	8	0.5, 0.5	0.3, 0.7
B	2918.1	2880	8	0.5, 0.5	0.2, 0.8
B	2918.7	2880	8	0.5, 0.5	0.1, 0.9
B	2912.2	2880	8	0.9, 0.1	0.9, 0.1
B	2912.3	2880	8	0.9, 0.1	0.8, 0.2
B	2912.5	2880	8	0.9, 0.1	0.7, 0.3
B	2912.9	2880	8	0.9, 0.1	0.6, 0.4
B	2913.5	2880	8	0.9, 0.1	0.5, 0.5
B	2914.3	2880	8	0.9, 0.1	0.4, 0.6
B	2915.3	2880	8	0.9, 0.1	0.3, 0.7
B	2916.5	2880	8	0.9, 0.1	0.2, 0.8
B	2917.9	2880	8	0.9, 0.1	0.1, 0.9
B	2928.4	2880	50	0.5, 0.5	0.9, 0.1
B	2926.5	2880	50	0.5, 0.5	0.8, 0.2
B	2924.7	2880	50	0.5, 0.5	0.7, 0.3
B	2923.2	2880	50	0.5, 0.5	0.6, 0.4
B	2922.6	2880	50	0.5, 0.5	0.5, 0.5
B	2923.4	2880	50	0.5, 0.5	0.4, 0.6

B	2925.7	2880	50	0.5, 0.5	0.3, 0.7
B	2929.2	2880	50	0.5, 0.5	0.2, 0.8
B	2933.5	2880	50	0.5, 0.5	0.1, 0.9
B	2916.9	2880	50	0.9, 0.1	0.9, 0.1
B	2917.0	2880	50	0.9, 0.1	0.8, 0.2
B	2917.9	2880	50	0.9, 0.1	0.7, 0.3
B	2920.1	2880	50	0.9, 0.1	0.6, 0.4
B	2923.7	2880	50	0.9, 0.1	0.5, 0.5
B	2928.7	2880	50	0.9, 0.1	0.4, 0.6
B	2935.1	2880	50	0.9, 0.1	0.3, 0.7
B	2941.1	2880	50	0.9, 0.1	0.2, 0.8
B	2948.1	2880	50	0.9, 0.1	0.1, 0.9
B	5831.0	5760	8	0.5, 0.5	0.9, 0.1
B	5829.8	5760	8	0.5, 0.5	0.8, 0.2
B	5828.9	5760	8	0.5, 0.5	0.7, 0.3
B	5828.4	5760	8	0.5, 0.5	0.6, 0.4
B	5828.1	5760	8	0.5, 0.5	0.5, 0.5
B	5828.3	5760	8	0.5, 0.5	0.4, 0.6
B	5828.8	5760	8	0.5, 0.5	0.3, 0.7
B	5829.7	5760	8	0.5, 0.5	0.2, 0.8
B	5831.1	5760	8	0.5, 0.5	0.1, 0.9
B	5821.8	5760	8	0.9, 0.1	0.9, 0.1
B	5822.0	5760	8	0.9, 0.1	0.8, 0.2
B	5822.6	5760	8	0.9, 0.1	0.7, 0.3
B	5823.5	5760	8	0.9, 0.1	0.6, 0.4
B	5824.7	5760	8	0.9, 0.1	0.5, 0.5
B	5826.3	5760	8	0.9, 0.1	0.4, 0.6
B	5828.4	5760	8	0.9, 0.1	0.3, 0.7
B	5830.8	5760	8	0.9, 0.1	0.2, 0.8
B	5833.6	5760	8	0.9, 0.1	0.1, 0.9
B	5854.8	5760	50	0.5, 0.5	0.9, 0.1
B	5851.1	5760	50	0.5, 0.5	0.8, 0.2
B	5847.3	5760	50	0.5, 0.5	0.7, 0.3
B	5844.2	5760	50	0.5, 0.5	0.6, 0.4
B	5842.8	5760	50	0.5, 0.5	0.5, 0.5
B	5844.3	5760	50	0.5, 0.5	0.4, 0.6
B	5848.7	5760	50	0.5, 0.5	0.3, 0.7
B	5855.8	5760	50	0.5, 0.5	0.2, 0.8
B	5864.5	5760	50	0.5, 0.5	0.1, 0.9

B	5827.9	5760	50	0.9, 0.1	0.9, 0.1
B	5828.4	5760	50	0.9, 0.1	0.8, 0.2
B	5830.9	5760	50	0.9, 0.1	0.7, 0.3
B	5836.2	5760	50	0.9, 0.1	0.6, 0.4
B	5845.2	5760	50	0.9, 0.1	0.5, 0.5
B	5857.8	5760	50	0.9, 0.1	0.4, 0.6
B	5873.3	5760	50	0.9, 0.1	0.3, 0.7
B	5890.4	5760	50	0.9, 0.1	0.2, 0.8
B	5907.5	5760	50	0.9, 0.1	0.1, 0.9
C	5838.0	5760	50	0.5, 0.5	0.9, 0.1
C	5827.8	5760	50	0.5, 0.5	0.8, 0.2
C	5819.7	5760	50	0.5, 0.5	0.7, 0.3
C	5814.5	5760	50	0.5, 0.5	0.6, 0.4
C	5812.7	5760	50	0.5, 0.5	0.5, 0.5
C	5814.4	5760	50	0.5, 0.5	0.4, 0.6
C	5818.8	5760	50	0.5, 0.5	0.3, 0.7
C	5823.4	5760	50	0.5, 0.5	0.2, 0.8
C	5829.6	5760	50	0.5, 0.5	0.1, 0.9
C	5821.6	5760	50	0.9, 0.1	0.9, 0.1
C	5823.1	5760	50	0.9, 0.1	0.8, 0.2
C	5828.2	5760	50	0.9, 0.1	0.7, 0.3
C	5837.4	5760	50	0.9, 0.1	0.6, 0.4
C	5851.5	5760	50	0.9, 0.1	0.5, 0.5
C	5869.7	5760	50	0.9, 0.1	0.4, 0.6
C	5887.5	5760	50	0.9, 0.1	0.3, 0.7
C	5900.5	5760	50	0.9, 0.1	0.2, 0.8
C	5908.8	5760	50	0.9, 0.1	0.1, 0.9
D	5842.0	5760	50	0.5, 0.5	0.9, 0.1
D	5838.0	5760	50	0.5, 0.5	0.8, 0.2
D	5833.5	5760	50	0.5, 0.5	0.7, 0.3
D	5829.8	5760	50	0.5, 0.5	0.6, 0.4
D	5828.4	5760	50	0.5, 0.5	0.5, 0.5
D	5829.8	5760	50	0.5, 0.5	0.4, 0.6
D	5833.5	5760	50	0.5, 0.5	0.3, 0.7
D	5837.9	5760	50	0.5, 0.5	0.2, 0.8
D	5841.9	5760	50	0.5, 0.5	0.1, 0.9
D	5822.7	5760	50	0.9, 0.1	0.9, 0.1
D	5823.2	5760	50	0.9, 0.1	0.8, 0.2

D	5825.6	5760	50	0.9, 0.1	0.7, 0.3
D	5831.3	5760	50	0.9, 0.1	0.6, 0.4
D	5841.5	5760	50	0.9, 0.1	0.5, 0.5
D	5855.8	5760	50	0.9, 0.1	0.4, 0.6
D	5871.1	5760	50	0.9, 0.1	0.3, 0.7
D	5883.6	5760	50	0.9, 0.1	0.2, 0.8
D	5892.1	5760	50	0.9, 0.1	0.1, 0.9
E	5816.2	5760	50	0.5, 0.5	0.9, 0.1
E	5811.6	5760	50	0.5, 0.5	0.8, 0.2
E	5807.9	5760	50	0.5, 0.5	0.7, 0.3
E	5805.3	5760	50	0.5, 0.5	0.6, 0.4
E	5804.4	5760	50	0.5, 0.5	0.5, 0.5
E	5805.4	5760	50	0.5, 0.5	0.4, 0.6
E	5807.7	5760	50	0.5, 0.5	0.3, 0.7
E	5810.6	5760	50	0.5, 0.5	0.2, 0.8
E	5813.6	5760	50	0.5, 0.5	0.1, 0.9
E	5817.5	5760	50	0.9, 0.1	0.9, 0.1
E	5818.4	5760	50	0.9, 0.1	0.8, 0.2
E	5821.4	5760	50	0.9, 0.1	0.7, 0.3
E	5827.4	5760	50	0.9, 0.1	0.6, 0.4
E	5837.0	5760	50	0.9, 0.1	0.5, 0.5
E	5849.1	5760	50	0.9, 0.1	0.4, 0.6
E	5860.0	5760	50	0.9, 0.1	0.3, 0.7
E	5867.1	5760	50	0.9, 0.1	0.2, 0.8
E	5870.9	5760	50	0.9, 0.1	0.1, 0.9
F	5821.1	5760	50	0.5, 0.5	0.9, 0.1
F	5820.2	5760	50	0.5, 0.5	0.8, 0.2
F	5819.8	5760	50	0.5, 0.5	0.7, 0.3
F	5819.7	5760	50	0.5, 0.5	0.6, 0.4
F	5819.7	5760	50	0.5, 0.5	0.5, 0.5
F	5819.7	5760	50	0.5, 0.5	0.4, 0.6
F	5819.8	5760	50	0.5, 0.5	0.3, 0.7
F	5820.2	5760	50	0.5, 0.5	0.2, 0.8
F	5821.1	5760	50	0.5, 0.5	0.1, 0.9
F	5823.5	5760	50	0.9, 0.1	0.9, 0.1
F	5823.7	5760	50	0.9, 0.1	0.8, 0.2
F	5824.3	5760	50	0.9, 0.1	0.7, 0.3
F	5824.8	5760	50	0.9, 0.1	0.6, 0.4
F	5825.1	5760	50	0.9, 0.1	0.5, 0.5

F	5824.8	5760	50	0.9, 0.1	0.4, 0.6
F	5824.2	5760	50	0.9, 0.1	0.3, 0.7
F	5823.7	5760	50	0.9, 0.1	0.2, 0.8
F	5823.5	5760	50	0.9, 0.1	0.1, 0.9

VI.10. Appendix J – Relative Percentage Bias for Multiple Membership Models

The percentage relative bias rates based on the posterior mean and the median are presented below for the scenarios considered in the main paper. This data shows that the estimator bias is practically equal for the posterior mean and median.

Percentage Relative Bias based on the Mean and Median Values for the Interviewer Variance Estimators

Change Profile Type	Mean	Median	N	% Change	W_{ij}	w_{ij}
A	-2.6	-5.1	2880	8	0.5, 0.5	0.9, 0.1
A	-0.1	-2.6	2880	8	0.5, 0.5	0.8, 0.2
A	1.8	-0.8	2880	8	0.5, 0.5	0.7, 0.3
A	2.8	0.2	2880	8	0.5, 0.5	0.6, 0.4
A	3.2	0.6	2880	8	0.5, 0.5	0.5, 0.5
A	2.8	0.2	2880	8	0.5, 0.5	0.4, 0.6
A	1.6	-1.0	2880	8	0.5, 0.5	0.3, 0.7
A	-0.3	-2.9	2880	8	0.5, 0.5	0.2, 0.8
A	-2.9	-5.4	2880	8	0.5, 0.5	0.1, 0.9
A	3.8	1.2	2880	8	0.9, 0.1	0.9, 0.1
A	4.8	2.2	2880	8	0.9, 0.1	0.8, 0.2
A	5.2	2.5	2880	8	0.9, 0.1	0.7, 0.3
A	4.7	2.1	2880	8	0.9, 0.1	0.6, 0.4
A	3.5	0.9	2880	8	0.9, 0.1	0.5, 0.5
A	1.5	-1.1	2880	8	0.9, 0.1	0.4, 0.6
A	-1.2	-3.7	2880	8	0.9, 0.1	0.3, 0.7
A	-4.5	-7.0	2880	8	0.9, 0.1	0.2, 0.8
A	-8.4	-10.8	2880	8	0.9, 0.1	0.1, 0.9
A	-35.4	-37.5	2880	50	0.5, 0.5	0.9, 0.1
A	-23.1	-25.4	2880	50	0.5, 0.5	0.8, 0.2
A	-12.1	-14.7	2880	50	0.5, 0.5	0.7, 0.3
A	-4.5	-7.3	2880	50	0.5, 0.5	0.6, 0.4
A	-1.9	-4.7	2880	50	0.5, 0.5	0.5, 0.5

A	-4.7	-7.5	2880	50	0.5, 0.5	0.4, 0.6
A	-12.4	-15.0	2880	50	0.5, 0.5	0.3, 0.7
A	-23.4	-25.7	2880	50	0.5, 0.5	0.2, 0.8
A	-35.8	-37.8	2880	50	0.5, 0.5	0.1, 0.9
A	-0.1	-2.7	2880	50	0.9, 0.1	0.9, 0.1
A	7.1	4.3	2880	50	0.9, 0.1	0.8, 0.2
A	9.7	6.8	2880	50	0.9, 0.1	0.7, 0.3
A	6.5	3.6	2880	50	0.9, 0.1	0.6, 0.4
A	-2.9	-5.6	2880	50	0.9, 0.1	0.5, 0.5
A	-17.1	-19.7	2880	50	0.9, 0.1	0.4, 0.6
A	-33.7	-36.1	2880	50	0.9, 0.1	0.3, 0.7
A	-50.3	-52.5	2880	50	0.9, 0.1	0.2, 0.8
A	-64.7	-66.8	2880	50	0.9, 0.1	0.1, 0.9
A	-4.8	-6.1	5760	8	0.5, 0.5	0.9, 0.1
A	-2.4	-3.6	5760	8	0.5, 0.5	0.8, 0.2
A	-0.6	-1.8	5760	8	0.5, 0.5	0.7, 0.3
A	0.5	-0.7	5760	8	0.5, 0.5	0.6, 0.4
A	0.9	-0.4	5760	8	0.5, 0.5	0.5, 0.5
A	0.5	-0.8	5760	8	0.5, 0.5	0.4, 0.6
A	-0.6	-1.9	5760	8	0.5, 0.5	0.3, 0.7
A	-2.4	-3.7	5760	8	0.5, 0.5	0.2, 0.8
A	-4.9	-6.1	5760	8	0.5, 0.5	0.1, 0.9
A	-2.4	-3.7	5760	8	0.7, 0.3	0.9, 0.1
A	-0.7	-2.0	5760	8	0.7, 0.3	0.8, 0.2
A	0.4	-0.9	5760	8	0.7, 0.3	0.7, 0.3
A	0.7	-0.6	5760	8	0.7, 0.3	0.6, 0.4
A	0.3	-1.0	5760	8	0.7, 0.3	0.5, 0.5
A	-0.9	-2.1	5760	8	0.7, 0.3	0.4, 0.6
A	-2.7	-4.0	5760	8	0.7, 0.3	0.3, 0.7
A	-5.2	-6.5	5760	8	0.7, 0.3	0.2, 0.8
A	-8.3	-9.5	5760	8	0.7, 0.3	0.1, 0.9
A	1.3	0.0	5760	8	0.9, 0.1	0.9, 0.1
A	2.3	1.0	5760	8	0.9, 0.1	0.8, 0.2
A	2.6	1.3	5760	8	0.9, 0.1	0.7, 0.3
A	2.2	0.9	5760	8	0.9, 0.1	0.6, 0.4
A	1.0	-0.3	5760	8	0.9, 0.1	0.5, 0.5
A	-0.9	-2.2	5760	8	0.9, 0.1	0.4, 0.6
A	-3.6	-4.8	5760	8	0.9, 0.1	0.3, 0.7
A	-6.8	-8.0	5760	8	0.9, 0.1	0.2, 0.8

A	-10.5	-11.7	5760	8	0.9, 0.1	0.1, 0.9
A	-13.3	-14.4	5760	21	0.5, 0.5	0.9, 0.1
A	-7.4	-8.6	5760	21	0.5, 0.5	0.8, 0.2
A	-2.9	-4.1	5760	21	0.5, 0.5	0.7, 0.3
A	0.1	-1.2	5760	21	0.5, 0.5	0.6, 0.4
A	1.1	-0.3	5760	21	0.5, 0.5	0.5, 0.5
A	0.0	-1.3	5760	21	0.5, 0.5	0.4, 0.6
A	-2.8	-4.1	5760	21	0.5, 0.5	0.3, 0.7
A	-7.4	-8.6	5760	21	0.5, 0.5	0.2, 0.8
A	-13.2	-14.4	5760	21	0.5, 0.5	0.1, 0.9
A	-21.7	-22.8	5760	33	0.5, 0.5	0.9, 0.1
A	-12.9	-14.1	5760	33	0.5, 0.5	0.8, 0.2
A	-5.6	-6.9	5760	33	0.5, 0.5	0.7, 0.3
A	-0.8	-2.2	5760	33	0.5, 0.5	0.6, 0.4
A	0.8	-0.6	5760	33	0.5, 0.5	0.5, 0.5
A	-1.0	-2.4	5760	33	0.5, 0.5	0.4, 0.6
A	-5.9	-7.2	5760	33	0.5, 0.5	0.3, 0.7
A	-13.3	-14.5	5760	33	0.5, 0.5	0.2, 0.8
A	-22.2	-23.3	5760	33	0.5, 0.5	0.1, 0.9
A	-32.9	-34.0	5760	50	0.5, 0.5	0.9, 0.1
A	-20.6	-21.8	5760	50	0.5, 0.5	0.8, 0.2
A	-9.7	-11.0	5760	50	0.5, 0.5	0.7, 0.3
A	-2.0	-3.4	5760	50	0.5, 0.5	0.6, 0.4
A	0.8	-0.7	5760	50	0.5, 0.5	0.5, 0.5
A	-2.0	-3.3	5760	50	0.5, 0.5	0.4, 0.6
A	-9.6	-10.9	5760	50	0.5, 0.5	0.3, 0.7
A	-20.6	-21.7	5760	50	0.5, 0.5	0.2, 0.8
A	-32.8	-33.9	5760	50	0.5, 0.5	0.1, 0.9
A	-16.6	-17.8	5760	50	0.7, 0.3	0.9, 0.1
A	-6.4	-7.7	5760	50	0.7, 0.3	0.8, 0.2
A	0.9	-0.4	5760	50	0.7, 0.3	0.7, 0.3
A	3.6	2.1	5760	50	0.7, 0.3	0.6, 0.4
A	0.6	-0.8	5760	50	0.7, 0.3	0.5, 0.5
A	-7.8	-9.1	5760	50	0.7, 0.3	0.4, 0.6
A	-19.9	-21.2	5760	50	0.7, 0.3	0.3, 0.7
A	-33.9	-35.0	5760	50	0.7, 0.3	0.2, 0.8
A	-47.7	-48.6	5760	50	0.7, 0.3	0.1, 0.9
A	1.2	-0.1	5760	50	0.9, 0.1	0.9, 0.1

A	8.4	7.0	5760	50	0.9, 0.1	0.8, 0.2
A	11.1	9.6	5760	50	0.9, 0.1	0.7, 0.3
A	8.0	6.6	5760	50	0.9, 0.1	0.6, 0.4
A	-1.1	-2.5	5760	50	0.9, 0.1	0.5, 0.5
A	-15.0	-16.3	5760	50	0.9, 0.1	0.4, 0.6
A	-31.3	-32.4	5760	50	0.9, 0.1	0.3, 0.7
A	-47.7	-48.6	5760	50	0.9, 0.1	0.2, 0.8
A	-62.5	-63.3	5760	50	0.9, 0.1	0.1, 0.9
A	-60.9	-61.9	5760	92	0.5, 0.5	0.9, 0.1
A	-43.8	-44.9	5760	92	0.5, 0.5	0.8, 0.2
A	-24.4	-25.7	5760	92	0.5, 0.5	0.7, 0.3
A	-7.9	-9.4	5760	92	0.5, 0.5	0.6, 0.4
A	-1.4	-3.0	5760	92	0.5, 0.5	0.5, 0.5
A	-8.3	-9.8	5760	92	0.5, 0.5	0.4, 0.6
A	-25.0	-26.4	5760	92	0.5, 0.5	0.3, 0.7
A	-44.5	-45.6	5760	92	0.5, 0.5	0.2, 0.8
A	-61.6	-62.5	5760	92	0.5, 0.5	0.1, 0.9
B	-7.20	-9.61	2880	8	0.5, 0.5	0.9, 0.1
B	-5.20	-7.66	2880	8	0.5, 0.5	0.8, 0.2
B	-3.32	-5.84	2880	8	0.5, 0.5	0.7, 0.3
B	-1.85	-4.41	2880	8	0.5, 0.5	0.6, 0.4
B	-0.72	-3.31	2880	8	0.5, 0.5	0.5, 0.5
B	-0.03	-2.64	2880	8	0.5, 0.5	0.4, 0.6
B	0.20	-2.40	2880	8	0.5, 0.5	0.3, 0.7
B	-0.07	-2.64	2880	8	0.5, 0.5	0.2, 0.8
B	-0.89	-3.44	2880	8	0.5, 0.5	0.1, 0.9
B	0.8	-1.8	2880	8	0.9, 0.1	0.9, 0.1
B	2.2	-0.4	2880	8	0.9, 0.1	0.8, 0.2
B	3.3	0.7	2880	8	0.9, 0.1	0.7, 0.3
B	4.1	1.5	2880	8	0.9, 0.1	0.6, 0.4
B	4.5	1.8	2880	8	0.9, 0.1	0.5, 0.5
B	4.4	1.7	2880	8	0.9, 0.1	0.4, 0.6
B	3.7	1.1	2880	8	0.9, 0.1	0.3, 0.7
B	2.7	0.0	2880	8	0.9, 0.1	0.2, 0.8
B	1.0	-1.6	2880	8	0.9, 0.1	0.1, 0.9
B	-39.2	-41.4	2880	50	0.5, 0.5	0.9, 0.1
B	-29.9	-32.3	2880	50	0.5, 0.5	0.8, 0.2
B	-19.6	-22.3	2880	50	0.5, 0.5	0.7, 0.3
B	-9.9	-12.8	2880	50	0.5, 0.5	0.6, 0.4

B	-3.5	-6.3	2880	50	0.5, 0.5	0.5, 0.5
B	-2.7	-5.4	2880	50	0.5, 0.5	0.4, 0.6
B	-8.0	-10.4	2880	50	0.5, 0.5	0.3, 0.7
B	-18.0	-20.2	2880	50	0.5, 0.5	0.2, 0.8
B	-30.8	-32.9	2880	50	0.5, 0.5	0.1, 0.9
B	-0.9	-3.4	2880	50	0.9, 0.1	0.9, 0.1
B	7.6	4.8	2880	50	0.9, 0.1	0.8, 0.2
B	15.0	12.0	2880	50	0.9, 0.1	0.7, 0.3
B	19.3	16.3	2880	50	0.9, 0.1	0.6, 0.4
B	18.6	15.8	2880	50	0.9, 0.1	0.5, 0.5
B	12.1	9.5	2880	50	0.9, 0.1	0.4, 0.6
B	0.6	-1.8	2880	50	0.9, 0.1	0.3, 0.7
B	-12.4	-14.6	2880	50	0.9, 0.1	0.2, 0.8
B	-28.5	-30.7	2880	50	0.9, 0.1	0.1, 0.9
B	-5.0	-6.2	5760	8	0.5, 0.5	0.9, 0.1
B	-3.0	-4.2	5760	8	0.5, 0.5	0.8, 0.2
B	-1.2	-2.4	5760	8	0.5, 0.5	0.7, 0.3
B	0.3	-0.9	5760	8	0.5, 0.5	0.6, 0.4
B	1.5	0.2	5760	8	0.5, 0.5	0.5, 0.5
B	2.2	0.9	5760	8	0.5, 0.5	0.4, 0.6
B	2.5	1.2	5760	8	0.5, 0.5	0.3, 0.7
B	2.2	0.9	5760	8	0.5, 0.5	0.2, 0.8
B	1.4	0.2	5760	8	0.5, 0.5	0.1, 0.9
B	0.4	-0.9	5760	8	0.9, 0.1	0.9, 0.1
B	1.7	0.5	5760	8	0.9, 0.1	0.8, 0.2
B	2.8	1.5	5760	8	0.9, 0.1	0.7, 0.3
B	3.6	2.2	5760	8	0.9, 0.1	0.6, 0.4
B	3.9	2.6	5760	8	0.9, 0.1	0.5, 0.5
B	3.8	2.5	5760	8	0.9, 0.1	0.4, 0.6
B	3.2	1.9	5760	8	0.9, 0.1	0.3, 0.7
B	2.1	0.8	5760	8	0.9, 0.1	0.2, 0.8
B	0.5	-0.8	5760	8	0.9, 0.1	0.1, 0.9
B	-37.2	-38.2	5760	50	0.5, 0.5	0.9, 0.1
B	-27.6	-28.8	5760	50	0.5, 0.5	0.8, 0.2
B	-17.1	-18.5	5760	50	0.5, 0.5	0.7, 0.3
B	-7.1	-8.5	5760	50	0.5, 0.5	0.6, 0.4
B	-0.3	-1.7	5760	50	0.5, 0.5	0.5, 0.5
B	0.9	-0.4	5760	50	0.5, 0.5	0.4, 0.6
B	-3.9	-5.1	5760	50	0.5, 0.5	0.3, 0.7

B	-13.6	-14.7	5760	50	0.5, 0.5	0.2, 0.8
B	-26.4	-27.3	5760	50	0.5, 0.5	0.1, 0.9
B	1.3	0.0	5760	50	0.9, 0.1	0.9, 0.1
B	11.5	10.1	5760	50	0.9, 0.1	0.8, 0.2
B	20.5	19.0	5760	50	0.9, 0.1	0.7, 0.3
B	25.9	24.3	5760	50	0.9, 0.1	0.6, 0.4
B	25.0	23.5	5760	50	0.9, 0.1	0.5, 0.5
B	17.2	15.8	5760	50	0.9, 0.1	0.4, 0.6
B	3.3	2.2	5760	50	0.9, 0.1	0.3, 0.7
B	-14.5	-15.5	5760	50	0.9, 0.1	0.2, 0.8
B	-34.1	-34.9	5760	50	0.9, 0.1	0.1, 0.9
C	-27.1	-28.2	5760	50	0.5, 0.5	0.9, 0.1
C	-15.8	-17.1	5760	50	0.5, 0.5	0.8, 0.2
C	-6.5	-7.9	5760	50	0.5, 0.5	0.7, 0.3
C	-0.6	-2.2	5760	50	0.5, 0.5	0.6, 0.4
C	0.5	-1.3	5760	50	0.5, 0.5	0.5, 0.5
C	-4.3	-6.2	5760	50	0.5, 0.5	0.4, 0.6
C	-14.2	-16.0	5760	50	0.5, 0.5	0.3, 0.7
C	-23.2	-24.9	5760	50	0.5, 0.5	0.2, 0.8
C	-36.1	-37.5	5760	50	0.5, 0.5	0.1, 0.9
C	0.6	-0.7	5760	50	0.9, 0.1	0.9, 0.1
C	8.6	7.1	5760	50	0.9, 0.1	0.8, 0.2
C	13.4	11.8	5760	50	0.9, 0.1	0.7, 0.3
C	12.8	11.0	5760	50	0.9, 0.1	0.6, 0.4
C	3.5	1.4	5760	50	0.9, 0.1	0.5, 0.5
C	-16.4	-18.7	5760	50	0.9, 0.1	0.4, 0.6
C	-39.1	-40.9	5760	50	0.9, 0.1	0.3, 0.7
C	-55.6	-56.9	5760	50	0.9, 0.1	0.2, 0.8
C	-65.9	-66.9	5760	50	0.9, 0.1	0.1, 0.9
D	-30.8	-31.9	5760	50	0.5, 0.5	0.9, 0.1
D	-21.6	-22.9	5760	50	0.5, 0.5	0.8, 0.2
D	-11.5	-12.9	5760	50	0.5, 0.5	0.7, 0.3
D	-2.9	-4.5	5760	50	0.5, 0.5	0.6, 0.4
D	0.5	-1.1	5760	50	0.5, 0.5	0.5, 0.5
D	-2.9	-4.4	5760	50	0.5, 0.5	0.4, 0.6
D	-11.5	-12.9	5760	50	0.5, 0.5	0.3, 0.7
D	-21.6	-22.8	5760	50	0.5, 0.5	0.2, 0.8
D	-30.8	-31.9	5760	50	0.5, 0.5	0.1, 0.9

D	0.1	-1.2	5760	50	0.9, 0.1	0.9, 0.1
D	10.4	8.9	5760	50	0.9, 0.1	0.8, 0.2
D	19.3	17.7	5760	50	0.9, 0.1	0.7, 0.3
D	23.3	21.6	5760	50	0.9, 0.1	0.6, 0.4
D	18.3	16.5	5760	50	0.9, 0.1	0.5, 0.5
D	3.3	1.6	5760	50	0.9, 0.1	0.4, 0.6
D	-16.4	-17.9	5760	50	0.9, 0.1	0.3, 0.7
D	-33.3	-34.4	5760	50	0.9, 0.1	0.2, 0.8
D	-45.0	-46.0	5760	50	0.9, 0.1	0.1, 0.9
E	-17.1	-18.2	5760	50	0.5, 0.5	0.9, 0.1
E	-9.4	-10.7	5760	50	0.5, 0.5	0.8, 0.2
E	-3.0	-4.5	5760	50	0.5, 0.5	0.7, 0.3
E	1.0	-0.7	5760	50	0.5, 0.5	0.6, 0.4
E	1.1	-0.8	5760	50	0.5, 0.5	0.5, 0.5
E	-3.5	-5.5	5760	50	0.5, 0.5	0.4, 0.6
E	-11.7	-13.5	5760	50	0.5, 0.5	0.3, 0.7
E	-20.9	-22.6	5760	50	0.5, 0.5	0.2, 0.8
E	-29.8	-31.3	5760	50	0.5, 0.5	0.1, 0.9
E	1.3	0.0	5760	50	0.9, 0.1	0.9, 0.1
E	9.9	8.4	5760	50	0.9, 0.1	0.8, 0.2
E	16.4	14.7	5760	50	0.9, 0.1	0.7, 0.3
E	17.9	15.8	5760	50	0.9, 0.1	0.6, 0.4
E	10.6	8.2	5760	50	0.9, 0.1	0.5, 0.5
E	-5.7	-8.2	5760	50	0.9, 0.1	0.4, 0.6
E	-23.4	-25.4	5760	50	0.9, 0.1	0.3, 0.7
E	-36.4	-38.0	5760	50	0.9, 0.1	0.2, 0.8
E	-45.2	-46.5	5760	50	0.9, 0.1	0.1, 0.9
F	-18.5	-19.7	5760	50	0.5, 0.5	0.9, 0.1
F	-11.7	-12.9	5760	50	0.5, 0.5	0.8, 0.2
F	-6.1	-7.5	5760	50	0.5, 0.5	0.7, 0.3
F	-2.4	-3.9	5760	50	0.5, 0.5	0.6, 0.4
F	-1.1	-2.6	5760	50	0.5, 0.5	0.5, 0.5
F	-2.4	-3.9	5760	50	0.5, 0.5	0.4, 0.6
F	-6.1	-7.4	5760	50	0.5, 0.5	0.3, 0.7
F	-11.7	-13.0	5760	50	0.5, 0.5	0.2, 0.8
F	-18.5	-19.7	5760	50	0.5, 0.5	0.1, 0.9
F	0.5	-0.8	5760	50	0.9, 0.1	0.9, 0.1
F	9.1	7.6	5760	50	0.9, 0.1	0.8, 0.2
F	16.2	14.6	5760	50	0.9, 0.1	0.7, 0.3

F	20.9	19.3	5760	50	0.9, 0.1	0.6, 0.4
F	22.5	20.8	5760	50	0.9, 0.1	0.5, 0.5
F	20.9	19.3	5760	50	0.9, 0.1	0.4, 0.6
F	16.2	14.6	5760	50	0.9, 0.1	0.3, 0.7
F	9.1	7.7	5760	50	0.9, 0.1	0.2, 0.8
F	0.6	-0.8	5760	50	0.9, 0.1	0.1, 0.9

VI.11. Appendix K – Confidence Interval Coverage Rates for Multiple Membership Models

The confidence interval coverage rates based on the Wald and the coverage rates based on the MCMC credible intervals are presented below for the scenarios considered in the main paper. Both measures show similar values and the same trends across factor changes. However, for some models with the most incorrect model weights the 95% credible interval performs just slightly better than the 95% Wald confidence interval, with an advantage of around 5%.

Wald and MCMC Credible Quantiles Confidence Interval Coverage Rates for the Interviewer Variance Estimator

Change Profile Type	Wald	MCMC	N	% Change	W_{ij}	w_{ij}
A	92.0	92.8	2880	8	0.5, 0.5	0.9, 0.1
A	92.4	93.3	2880	8	0.5, 0.5	0.8, 0.2
A	93.2	93.3	2880	8	0.5, 0.5	0.7, 0.3
A	93.6	93.6	2880	8	0.5, 0.5	0.6, 0.4
A	93.6	93.6	2880	8	0.5, 0.5	0.5, 0.5
A	93.8	93.7	2880	8	0.5, 0.5	0.4, 0.6
A	93.3	93.9	2880	8	0.5, 0.5	0.3, 0.7
A	92.9	92.8	2880	8	0.5, 0.5	0.2, 0.8
A	92.1	92.3	2880	8	0.5, 0.5	0.1, 0.9
A	96.4	95.6	2880	8	0.9, 0.1	0.9, 0.1
A	96.4	95.4	2880	8	0.9, 0.1	0.8, 0.2
A	96.5	94.9	2880	8	0.9, 0.1	0.7, 0.3
A	96.6	94.9	2880	8	0.9, 0.1	0.6, 0.4
A	96.4	94.8	2880	8	0.9, 0.1	0.5, 0.5
A	95.7	95.1	2880	8	0.9, 0.1	0.4, 0.6
A	94.9	95.5	2880	8	0.9, 0.1	0.3, 0.7
A	92.9	94.6	2880	8	0.9, 0.1	0.2, 0.8
A	91.1	93.6	2880	8	0.9, 0.1	0.1, 0.9
A	61.3	68.4	2880	50	0.5, 0.5	0.9, 0.1
A	79.5	82.7	2880	50	0.5, 0.5	0.8, 0.2

A	89.1	90.9	2880	50	0.5, 0.5	0.7, 0.3
A	92.3	94.2	2880	50	0.5, 0.5	0.6, 0.4
A	93.9	94.3	2880	50	0.5, 0.5	0.5, 0.5
A	93.7	94.5	2880	50	0.5, 0.5	0.4, 0.6
A	89.0	91.9	2880	50	0.5, 0.5	0.3, 0.7
A	79.6	84.2	2880	50	0.5, 0.5	0.2, 0.8
A	60.8	68.3	2880	50	0.5, 0.5	0.1, 0.9
A	93.7	93.5	2880	50	0.9, 0.1	0.9, 0.1
A	94.9	93.5	2880	50	0.9, 0.1	0.8, 0.2
A	95.0	93.5	2880	50	0.9, 0.1	0.7, 0.3
A	94.6	93.8	2880	50	0.9, 0.1	0.6, 0.4
A	93.0	93.0	2880	50	0.9, 0.1	0.5, 0.5
A	85.7	88.2	2880	50	0.9, 0.1	0.4, 0.6
A	67.7	73.0	2880	50	0.9, 0.1	0.3, 0.7
A	39.2	46.1	2880	50	0.9, 0.1	0.2, 0.8
A	12.6	18.1	2880	50	0.9, 0.1	0.1, 0.9
A	92.1	93.5	5760	8	0.5, 0.5	0.9, 0.1
A	93.8	94.5	5760	8	0.5, 0.5	0.8, 0.2
A	94.1	94.6	5760	8	0.5, 0.5	0.7, 0.3
A	94.3	94.8	5760	8	0.5, 0.5	0.6, 0.4
A	94.6	94.8	5760	8	0.5, 0.5	0.5, 0.5
A	94.4	94.8	5760	8	0.5, 0.5	0.4, 0.6
A	93.7	94.6	5760	8	0.5, 0.5	0.3, 0.7
A	93.4	93.8	5760	8	0.5, 0.5	0.2, 0.8
A	92.1	93.3	5760	8	0.5, 0.5	0.1, 0.9
A	93.7	94.9	5760	8	0.7, 0.3	0.9, 0.1
A	94.4	94.8	5760	8	0.7, 0.3	0.8, 0.2
A	95.0	95.2	5760	8	0.7, 0.3	0.7, 0.3
A	94.8	95.2	5760	8	0.7, 0.3	0.6, 0.4
A	94.8	95.1	5760	8	0.7, 0.3	0.5, 0.5
A	94.4	94.7	5760	8	0.7, 0.3	0.4, 0.6
A	93.4	94.4	5760	8	0.7, 0.3	0.3, 0.7
A	92.3	93.5	5760	8	0.7, 0.3	0.2, 0.8
A	89.4	91.3	5760	8	0.7, 0.3	0.1, 0.9
A	94.5	94.3	5760	8	0.9, 0.1	0.9, 0.1
A	94.6	94.3	5760	8	0.9, 0.1	0.8, 0.2
A	94.6	94.1	5760	8	0.9, 0.1	0.7, 0.3
A	94.4	93.8	5760	8	0.9, 0.1	0.6, 0.4
A	93.7	94.3	5760	8	0.9, 0.1	0.5, 0.5

A	93.6	94.0	5760	8	0.9, 0.1	0.4, 0.6
A	92.4	93.4	5760	8	0.9, 0.1	0.3, 0.7
A	90.4	91.7	5760	8	0.9, 0.1	0.2, 0.8
A	87.5	89.4	5760	8	0.9, 0.1	0.1, 0.9
A	81.6	84.7	5760	21	0.5, 0.5	0.9, 0.1
A	88.1	90.6	5760	21	0.5, 0.5	0.8, 0.2
A	92.0	93.3	5760	21	0.5, 0.5	0.7, 0.3
A	93.2	94.5	5760	21	0.5, 0.5	0.6, 0.4
A	94.2	94.5	5760	21	0.5, 0.5	0.5, 0.5
A	93.7	94.1	5760	21	0.5, 0.5	0.4, 0.6
A	91.8	93.2	5760	21	0.5, 0.5	0.3, 0.7
A	88.3	90.9	5760	21	0.5, 0.5	0.2, 0.8
A	83.5	85.8	5760	21	0.5, 0.5	0.1, 0.9
A	71.0	75.7	5760	33	0.5, 0.5	0.9, 0.1
A	85.7	89.1	5760	33	0.5, 0.5	0.8, 0.2
A	93.0	94.1	5760	33	0.5, 0.5	0.7, 0.3
A	94.7	95.4	5760	33	0.5, 0.5	0.6, 0.4
A	95.2	96.0	5760	33	0.5, 0.5	0.5, 0.5
A	94.4	95.9	5760	33	0.5, 0.5	0.4, 0.6
A	91.9	93.5	5760	33	0.5, 0.5	0.3, 0.7
A	84.7	87.3	5760	33	0.5, 0.5	0.2, 0.8
A	70.7	74.8	5760	33	0.5, 0.5	0.1, 0.9
A	44.1	51.0	5760	50	0.5, 0.5	0.9, 0.1
A	74.5	78.5	5760	50	0.5, 0.5	0.8, 0.2
A	88.6	90.7	5760	50	0.5, 0.5	0.7, 0.3
A	93.5	93.9	5760	50	0.5, 0.5	0.6, 0.4
A	94.7	94.8	5760	50	0.5, 0.5	0.5, 0.5
A	94.4	94.4	5760	50	0.5, 0.5	0.4, 0.6
A	89.9	92.0	5760	50	0.5, 0.5	0.3, 0.7
A	73.5	78.6	5760	50	0.5, 0.5	0.2, 0.8
A	45.1	50.8	5760	50	0.5, 0.5	0.1, 0.9
A	79.1	82.8	5760	50	0.7, 0.3	0.9, 0.1
A	90.4	91.8	5760	50	0.7, 0.3	0.8, 0.2
A	94.1	93.3	5760	50	0.7, 0.3	0.7, 0.3
A	94.4	93.8	5760	50	0.7, 0.3	0.6, 0.4
A	94.3	93.8	5760	50	0.7, 0.3	0.5, 0.5
A	90.8	92.0	5760	50	0.7, 0.3	0.4, 0.6
A	76.2	80.2	5760	50	0.7, 0.3	0.3, 0.7
A	45.4	50.6	5760	50	0.7, 0.3	0.2, 0.8
A	12.9	18.1	5760	50	0.7, 0.3	0.1, 0.9

A	95.2	95.0	5760	50	0.9, 0.1	0.9, 0.1
A	95.4	93.7	5760	50	0.9, 0.1	0.8, 0.2
A	95.5	92.3	5760	50	0.9, 0.1	0.7, 0.3
A	95.6	93.9	5760	50	0.9, 0.1	0.6, 0.4
A	94.6	94.9	5760	50	0.9, 0.1	0.5, 0.5
A	83.4	86.5	5760	50	0.9, 0.1	0.4, 0.6
A	54.4	59.6	5760	50	0.9, 0.1	0.3, 0.7
A	15.7	20.3	5760	50	0.9, 0.1	0.2, 0.8
A	1.7	2.3	5760	50	0.9, 0.1	0.1, 0.9
A	4.4	5.7	5760	92	0.5, 0.5	0.9, 0.1
A	33.1	38.7	5760	92	0.5, 0.5	0.8, 0.2
A	73.1	77.1	5760	92	0.5, 0.5	0.7, 0.3
A	91.1	91.8	5760	92	0.5, 0.5	0.6, 0.4
A	92.7	92.8	5760	92	0.5, 0.5	0.5, 0.5
A	90.0	91.2	5760	92	0.5, 0.5	0.4, 0.6
A	72.6	76.8	5760	92	0.5, 0.5	0.3, 0.7
A	30.6	35.0	5760	92	0.5, 0.5	0.2, 0.8
A	3.9	5.0	5760	92	0.5, 0.5	0.1, 0.9
B	90.60	92.90	2880	8	0.5, 0.5	0.9, 0.1
B	92.30	93.40	2880	8	0.5, 0.5	0.8, 0.2
B	93.20	94.00	2880	8	0.5, 0.5	0.7, 0.3
B	93.60	94.00	2880	8	0.5, 0.5	0.6, 0.4
B	94.00	94.20	2880	8	0.5, 0.5	0.5, 0.5
B	94.10	94.40	2880	8	0.5, 0.5	0.4, 0.6
B	93.90	94.20	2880	8	0.5, 0.5	0.3, 0.7
B	94.00	94.50	2880	8	0.5, 0.5	0.2, 0.8
B	93.70	94.40	2880	8	0.5, 0.5	0.1, 0.9
B	94.9	94.7	2880	8	0.9, 0.1	0.9, 0.1
B	95.6	94.4	2880	8	0.9, 0.1	0.8, 0.2
B	95.4	94.4	2880	8	0.9, 0.1	0.7, 0.3
B	95.5	94.2	2880	8	0.9, 0.1	0.6, 0.4
B	95.4	94.1	2880	8	0.9, 0.1	0.5, 0.5
B	95.7	94.4	2880	8	0.9, 0.1	0.4, 0.6
B	95.4	94.5	2880	8	0.9, 0.1	0.3, 0.7
B	95.0	94.9	2880	8	0.9, 0.1	0.2, 0.8
B	95.0	94.9	2880	8	0.9, 0.1	0.1, 0.9
B	57.6	63.8	2880	50	0.5, 0.5	0.9, 0.1
B	71.8	77.7	2880	50	0.5, 0.5	0.8, 0.2

B	82.7	86.2	2880	50	0.5, 0.5	0.7, 0.3
B	88.8	90.7	2880	50	0.5, 0.5	0.6, 0.4
B	90.9	93.2	2880	50	0.5, 0.5	0.5, 0.5
B	91.6	93.0	2880	50	0.5, 0.5	0.4, 0.6
B	90.6	91.1	2880	50	0.5, 0.5	0.3, 0.7
B	84.9	87.3	2880	50	0.5, 0.5	0.2, 0.8
B	72.5	76.9	2880	50	0.5, 0.5	0.1, 0.9
B	92.8	93.5	2880	50	0.9, 0.1	0.9, 0.1
B	94.3	93.1	2880	50	0.9, 0.1	0.8, 0.2
B	94.3	91.1	2880	50	0.9, 0.1	0.7, 0.3
B	93.6	89.7	2880	50	0.9, 0.1	0.6, 0.4
B	93.6	90.3	2880	50	0.9, 0.1	0.5, 0.5
B	93.8	92.1	2880	50	0.9, 0.1	0.4, 0.6
B	93.7	93.1	2880	50	0.9, 0.1	0.3, 0.7
B	86.5	88.2	2880	50	0.9, 0.1	0.2, 0.8
B	73.0	76.1	2880	50	0.9, 0.1	0.1, 0.9
B	90.6	93.1	5760	8	0.5, 0.5	0.9, 0.1
B	92.5	94.1	5760	8	0.5, 0.5	0.8, 0.2
B	93.6	94.5	5760	8	0.5, 0.5	0.7, 0.3
B	94.1	94.6	5760	8	0.5, 0.5	0.6, 0.4
B	94.5	94.7	5760	8	0.5, 0.5	0.5, 0.5
B	94.7	95.1	5760	8	0.5, 0.5	0.4, 0.6
B	94.6	94.8	5760	8	0.5, 0.5	0.3, 0.7
B	94.8	95.0	5760	8	0.5, 0.5	0.2, 0.8
B	94.5	94.7	5760	8	0.5, 0.5	0.1, 0.9
B	94.3	95.0	5760	8	0.9, 0.1	0.9, 0.1
B	94.7	94.9	5760	8	0.9, 0.1	0.8, 0.2
B	95.1	95.0	5760	8	0.9, 0.1	0.7, 0.3
B	95.5	94.9	5760	8	0.9, 0.1	0.6, 0.4
B	95.3	95.1	5760	8	0.9, 0.1	0.5, 0.5
B	95.5	95.1	5760	8	0.9, 0.1	0.4, 0.6
B	95.1	95.0	5760	8	0.9, 0.1	0.3, 0.7
B	94.8	95.1	5760	8	0.9, 0.1	0.2, 0.8
B	94.6	94.8	5760	8	0.9, 0.1	0.1, 0.9
B	34.5	40.0	5760	50	0.5, 0.5	0.9, 0.1
B	60.2	65.4	5760	50	0.5, 0.5	0.8, 0.2
B	80.4	85.0	5760	50	0.5, 0.5	0.7, 0.3
B	91.9	93.2	5760	50	0.5, 0.5	0.6, 0.4
B	94.4	93.7	5760	50	0.5, 0.5	0.5, 0.5

B	94.3	93.5	5760	50	0.5, 0.5	0.4, 0.6
B	93.3	93.8	5760	50	0.5, 0.5	0.3, 0.7
B	86.2	88.3	5760	50	0.5, 0.5	0.2, 0.8
B	66.0	69.7	5760	50	0.5, 0.5	0.1, 0.9
B	95.9	95.7	5760	50	0.9, 0.1	0.9, 0.1
B	96.1	93.6	5760	50	0.9, 0.1	0.8, 0.2
B	91.7	86.4	5760	50	0.9, 0.1	0.7, 0.3
B	87.0	80.3	5760	50	0.9, 0.1	0.6, 0.4
B	87.9	81.8	5760	50	0.9, 0.1	0.5, 0.5
B	92.9	89.5	5760	50	0.9, 0.1	0.4, 0.6
B	95.0	95.6	5760	50	0.9, 0.1	0.3, 0.7
B	85.2	87.9	5760	50	0.9, 0.1	0.2, 0.8
B	50.6	55.2	5760	50	0.9, 0.1	0.1, 0.9
C	57.7	62.9	5760	50	0.5, 0.5	0.9, 0.1
C	79.0	82.5	5760	50	0.5, 0.5	0.8, 0.2
C	89.8	91.4	5760	50	0.5, 0.5	0.7, 0.3
C	93.0	94.0	5760	50	0.5, 0.5	0.6, 0.4
C	93.2	94.3	5760	50	0.5, 0.5	0.5, 0.5
C	90.9	92.7	5760	50	0.5, 0.5	0.4, 0.6
C	81.7	85.9	5760	50	0.5, 0.5	0.3, 0.7
C	66.1	71.6	5760	50	0.5, 0.5	0.2, 0.8
C	37.3	43.2	5760	50	0.5, 0.5	0.1, 0.9
C	94.6	94.7	5760	50	0.9, 0.1	0.9, 0.1
C	94.9	93.9	5760	50	0.9, 0.1	0.8, 0.2
C	94.4	91.7	5760	50	0.9, 0.1	0.7, 0.3
C	94.4	92.0	5760	50	0.9, 0.1	0.6, 0.4
C	93.6	93.7	5760	50	0.9, 0.1	0.5, 0.5
C	82.8	85.7	5760	50	0.9, 0.1	0.4, 0.6
C	41.7	47.9	5760	50	0.9, 0.1	0.3, 0.7
C	6.1	9.8	5760	50	0.9, 0.1	0.2, 0.8
C	0.3	0.8	5760	50	0.9, 0.1	0.1, 0.9
D	49.3	55.1	5760	50	0.5, 0.5	0.9, 0.1
D	72.4	77.7	5760	50	0.5, 0.5	0.8, 0.2
D	87.6	89.8	5760	50	0.5, 0.5	0.7, 0.3
D	93.1	93.1	5760	50	0.5, 0.5	0.6, 0.4
D	93.6	93.5	5760	50	0.5, 0.5	0.5, 0.5
D	92.1	92.9	5760	50	0.5, 0.5	0.4, 0.6
D	86.5	88.8	5760	50	0.5, 0.5	0.3, 0.7
D	72.1	76.6	5760	50	0.5, 0.5	0.2, 0.8

D	50.2	56.3	5760	50	0.5, 0.5	0.1, 0.9
D	93.8	94.2	5760	50	0.9, 0.1	0.9, 0.1
D	94.1	91.9	5760	50	0.9, 0.1	0.8, 0.2
D	90.8	87.2	5760	50	0.9, 0.1	0.7, 0.3
D	89.4	85.1	5760	50	0.9, 0.1	0.6, 0.4
D	91.7	88.7	5760	50	0.9, 0.1	0.5, 0.5
D	94.6	94.2	5760	50	0.9, 0.1	0.4, 0.6
D	81.1	83.9	5760	50	0.9, 0.1	0.3, 0.7
D	47.5	54.2	5760	50	0.9, 0.1	0.2, 0.8
D	19.1	24.3	5760	50	0.9, 0.1	0.1, 0.9
E	76.9	81.3	5760	50	0.5, 0.5	0.9, 0.1
E	88.2	90.4	5760	50	0.5, 0.5	0.8, 0.2
E	93.0	94.1	5760	50	0.5, 0.5	0.7, 0.3
E	94.8	94.9	5760	50	0.5, 0.5	0.6, 0.4
E	94.7	95.1	5760	50	0.5, 0.5	0.5, 0.5
E	93.2	94.0	5760	50	0.5, 0.5	0.4, 0.6
E	85.1	90.0	5760	50	0.5, 0.5	0.3, 0.7
E	73.9	79.1	5760	50	0.5, 0.5	0.2, 0.8
E	54.0	60.9	5760	50	0.5, 0.5	0.1, 0.9
E	94.4	93.5	5760	50	0.9, 0.1	0.9, 0.1
E	94.4	92.3	5760	50	0.9, 0.1	0.8, 0.2
E	92.2	89.1	5760	50	0.9, 0.1	0.7, 0.3
E	92.5	89.0	5760	50	0.9, 0.1	0.6, 0.4
E	93.0	91.6	5760	50	0.9, 0.1	0.5, 0.5
E	89.2	92.5	5760	50	0.9, 0.1	0.4, 0.6
E	69.3	75.6	5760	50	0.9, 0.1	0.3, 0.7
E	42.2	49.7	5760	50	0.9, 0.1	0.2, 0.8
E	16.9	22.4	5760	50	0.9, 0.1	0.1, 0.9
F	73.3	78.3	5760	50	0.5, 0.5	0.9, 0.1
F	85.1	87.5	5760	50	0.5, 0.5	0.8, 0.2
F	90.0	91.5	5760	50	0.5, 0.5	0.7, 0.3
F	92.0	93.3	5760	50	0.5, 0.5	0.6, 0.4
F	92.7	93.4	5760	50	0.5, 0.5	0.5, 0.5
F	92.2	93.1	5760	50	0.5, 0.5	0.4, 0.6
F	90.1	91.5	5760	50	0.5, 0.5	0.3, 0.7
F	84.9	87.2	5760	50	0.5, 0.5	0.2, 0.8
F	73.3	78.3	5760	50	0.5, 0.5	0.1, 0.9
F	94.2	94.4	5760	50	0.9, 0.1	0.9, 0.1

F	94.8	93.0	5760	50	0.9, 0.1	0.8, 0.2
F	93.0	89.8	5760	50	0.9, 0.1	0.7, 0.3
F	90.8	86.4	5760	50	0.9, 0.1	0.6, 0.4
F	90.0	84.5	5760	50	0.9, 0.1	0.5, 0.5
F	90.8	86.0	5760	50	0.9, 0.1	0.4, 0.6
F	93.2	89.5	5760	50	0.9, 0.1	0.3, 0.7
F	94.9	93.2	5760	50	0.9, 0.1	0.2, 0.8
F	94.3	94.3	5760	50	0.9, 0.1	0.1, 0.9

VI.12. Appendix L – Distribution of Interviewers and Areas in the FACS Dataset

These tables present the distribution of the number of interviewers per area and the distribution of the number of areas per interviewer respectively for the dataset analysed in Paper 1.

Table showing the Distribution of Number of Interviewers per Area

Interviewers per Area	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Number of Areas	0	1	14	13	20	22	25	21	12	12	4	4	1	0	1	150
% Areas	0	0.7	9.3	8.7	13.3	14.7	16.7	14.0	8.0	8.0	2.7	2.7	0.7	0.0	0.7	100

Table showing the Distribution of Number of Areas per Interviewer

Areas per Interviewer	1	2	3	4	5	6	7	8	9	10	Total
Number of Interviewers	61	99	72	39	23	27	6	4	3	1	335
% Interviewers	18.2	29.6	21.5	11.6	6.9	8.1	1.8	1.2	0.9	0.3	100

VII. Reference List

Anderson, D. A. & Aitken, M. (1985). Variance component models with binary response: interviewer variability. *Journal of Royal Statistical Society, B, 47,* 203-210.

Austin, P.C. (2005). Bias in penalized quasi-likelihood estimation in random effects logistic regression models when the random effects are not normally distributed. *Communications in Statistics—Simulation and Computation*, *34*, 549-565.

Benet-Martínez, V. & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75,* 729-750.

Berg, A., Meyer, R. & Yu., J. (2004). Deviance Information Criterion for Comparing Stochastic Volatility Models. *Journal of Business and Economic Statistics, 22,* 107-20.

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. J. (1998). Applying MCMC Methods to Multi-level Models. PhD thesis, University of Bath.

Browne, W. J. (2012). *MCMC estimation in MLwiN. Version 2.25.* Bristol, Centre for Multilevel Modelling.

Browne, W. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1,* 473-514.

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.

Campanelli, P. & O'Muircheartaigh, C. (2002). The importance of experimental control in testing the impact of interviewer continuity on panel survey nonresponse. *Quality & Quantity, 36,* 129–144.

Campanelli, P., Sturgis, P. & Purdon, S. (1997). *Can you hear me knocking: An investigation into the impact of interviewers on survey response rates*. Final Report for UK ESRC Grant R000235776, London, National Centre for Social Research.

Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis 4,* 651–674.

Chandola, T., Clarke, P., Wiggins, R. D. & Bartley, M. (2005). Who you live with and where you live: Setting the context for health using multiple membership multilevel models. *Journal of Epidemiology & Community Health, 59(2),* 170–175.

Chung, H. & Beretvas, S. N. (2012). The Impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology, 65,* 185–200.

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

De Leeuw, E.D., Hox, J.J., Snijkers, G. & De Heer, W. (1998). Interviewer opinions, attitudes, and strategies regarding survey participation and their effect on response. *ZUMA-Nachrichten Spezial, 4,* 239-248.

Department for Work and Pensions (n.d.). *Families and Children Study. Sample design and response*. Retrieved January 27, 2011 from http://research.dwp.gov.uk/asd/asd5/facs/facs_sample.asp.

Dundas, R., Leyland, A. H., Macintyre, S. & Leon, D. A. (2006). Does the primary school attended influence self-reported health or its risk factors in later life? Aberdeen Children of the 1950s Study. *International Journal of Epidemiology, 35,* 458-465

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. (2002). Teaching groups as foci for evaluating performance in cost-effectiveness of GCE Advanced Level provision: Some practical methodological innovations. *School Effectiveness and School Improvement, 13,* 225-246.

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.

Fielding, A., Thomas, H. F. S., Steele, F., Browne, W., Leyland, A., Spencer, N. & Davison, I. (2006b). *Using Cross-classified Multilevel Models to Improve Estimates of the Determination of Pupil Attainment: A Scoping Study.* Research Report for the Department for Education and Skills. School of Education, University of Birmingham. Retrieved August 8, 2013 from http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&cad=rja&ved=0CC0QFjAA&url=http%3A%2F%2Fwww.bristol.ac.uk%2Fcmm%2Fteam%2Fdfes-scoping-report.pdf&ei=V48DUsGoFMmd0wXuioGIBQ&usg=AFQjCNH1kBp84MuMfjdqsgGJGm8elv0xlg&sig2=HNuujnoilLDbh0WeSJ6zWw

Fielding, A. & Yang, M. (2005). Generalized linear mixed models for ordered responses in complex multilevel structures: effects beneath the school or college in education. *Journal of the Royal Statistical Society: Series A, 168,* 159–183.

Ferede, T. (2013). Multilevel modelling of modern contraceptive use among rural and urban population of Ethiopia. *American Journal of Mathematics and Statistics, 3(1),* 1–16.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin D. B. (2004). *Bayesian Data Analysis.* Second Edition. Chapman & Hall, London.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research, 22,* 364–375.

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

Goldstein, H. (2011b). Estimating research performance by using research grant award gradings. *Journal of the Royal Statistical Society: Series A, 174,* 83–93.

Kizilkaya, K. & Tempelman, R.J. (2003). Cumulative t-link threshold models for genetic analysis of calving ease scores. *Genet. Sel. Evol. 35,* 489–512.

Goldstein, H., Browne, W. J. & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics 1,* 223–231.

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, A, 159,* 505–13.

Groves, R.M., Cialdini, R.B. & Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56(4),* 475–495.

Groves, R.M. & Couper, M.P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Hansen, K. M. (2006). The effects of incentives, interview length, and interviewer characteristics on response rates in a CATI study. *International Journal of Public Opinion Research, 19,* 112–121.

Hansen, M., Hurwitz, W. & Madow, W. (1953). *Sample Surveys Methods and Theory. New* York: John Wiley and Sons.

Hansen, M., Hurwitz, W. & Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute, 38,* 359–374.

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

Hill, D. H. & Willis, R. J. (2001). Reducing panel attrition: A search for effective policy instruments. *The Journal of Human Resources, 36(3),* 416-438.

Hox, J. & De Leeuw, E. (2002). The influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little, R. J. A. *Survey Nonresponse* (pp. 103-119). New York: Wiley.

Johnell, K., Merlo, J., Lynch, J. & Blennow, G. (2004). Neighbourhood social participation and women's use of anxiolytic-hypnotic drugs: a multilevel analysis. *Journal of Epidemiology & Community Health*, 58, 59-64.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association, 57,* 92-115.

Laurie, H., Smith, R. & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics, 15(2),* 269-282.

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

Luo, W. & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research, 44(2),* 182-212.

Lynn, P., Kaminska, O. & Goldstein, H. (2013). *Panel attrition: How important is it to keep the same interviewer?* Manuscript submitted for publication.

Lyon, N., Mangla, J., Tait, C. & Scholes, S. (2007). *Families and Children Study (FACS) 2005, Wave 7 Technical Report.* National Centre for Social Research: London.

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.

Meyers, J. L. & Beretvas, S. N. (2006). The impact of inappropriate modeling of crossclassified data structures. *Multivariate Behavioral Research, 41(4),* 473-497.

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.

Moorman, P. G., Newman, B., Millikan, R. C., Tse, C. J. & Sandler, D. P. (1999). Participation rates in a case-control study: The impact of age, race, and race of interviewer. *Annals of Epidemiology, 9(3),* 188-95.

Morton-Williams, J. (1993). *Interviewer Approaches.* Aldershot: Dartmouth Publishing Company Limited.

Nicoletti, C. & Buck, N. (2004). Explaining interviewee contact and co-operation in the British and German Household Panel. *ISER Working paper Series, 2004-06, Institute for Social & Economic Research, ESRC.*

Nicoletti, C. & Peracchi, F. (2005). Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, Series A, 168(4),* 763-781.

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.

Rasbash, J. (2006). *Why use multilevel modelling?* Retrieved August 2, 2010 from University of Bristol, Centre for Multilevel Modelling Web site: http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/jon-rasbash.ppt

Rasbash, J., Steele, F., Browne, W. J. & Goldstein, H. (2012). *A user's guide to MLwiN, v2.26.* Centre for Multilevel Modelling, University of Bristol.

Ritter, F., Schoelles, M. J., Quigley, K. S. & Klein, L. C. (1996). Determining the number of simulation runs: Treating simulations as theories by not sampling their behavior. In Narayanan, S. & Rothrock, L. (Eds.), *Human-in-the-loop Simulations: Methods and Practice* (pp. 97-116). London: Springer-Verlag.

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

Schatteman, T. (2000). *Do interviewer characteristics influence respondents' participation in panel surveys?* Paper presented at 2004 JRSS/ESRC conference on Statistical methods for attrition and nonresponse in social surveys, London.

Scherbaum, C. A. & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12,* 347-367.

Sinibaldi, J., Jäckle, A., Tipping, S. & Lynn, P. (2009). *Interviewer characteristics, their doorstep behavior, and survey co-operation.* Proceedings of the Survey Research Methods Section of the 2009 Joint Statistical Meetings.

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.

Snijkers, G., Hox, J. J. & De Leeuw, E.D. (1999). Interviewers' tactics for fighting survey nonresponse. *Journal of Official Statistics, 15(2),* 185-198.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64(4),* 583-640.

Steele, F., Clarke, P. & Washbrook, E. (2013). Modeling household decisions using longitudinal data from household panel surveys, with applications to residential mobility. *Sociological Methodology, 43(1),* 225-276.

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.

Tranmer, M. & Steel, D. G. (2001). Ignoring a level in a multilevel model: evidence from UK census data. *Environment and Planning A, 33(5),* 941 - 948.

van den Noortgate, W., Opdenakker, M. & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement, 16(3),* 281–303.

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/.

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling, 211,* 1–10.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review, 19(4),* 361–378.

Watson, N. & Wooden, M. (2006). *Modelling Longitudinal Survey Response: The Experience of the HILDA Survey.* Paper presented at the ASCPRI Social Science Methodology Conference.

Wilberg, M. J. & Bence, J. R. (2008). Performance of deviance information criterion model selection in statistical catch-at-age analysis. *Fisheries Research, 93,* 212–221.

Weinhardt, M. & Kreuter, F. (2011). *The different roles of interviewers: How does interviewer personality affect respondents' survey participation and response behavior?* Working paper.

Zhu, L. & Carlin, B. P. (2000). Comparing Hierarchical Models for Spatio-Temporally Misaligned Data Using the Deviance Information Criterion. *Statistics in Medicine, 19,* 2265–2278.