Assessment of Multiple Membership Multilevel Models: An Application to Interviewer Effects on Nonresponse

Gabriele B. Durrant, Rebecca Vassallo and Peter W.F. Smith

Department of Social Statistics and Demography

University of Southampton, UK

Address for Correspondence:
Gabriele B. Durrant
Department of Social Statistics and Demography
University of Southampton
SO17 1 BJ Southampton
United Kingdom
g.durrant@southampton.ac.uk

Acknowledgements
Durrant's and Smith's work was supported by ESRC grant number ES/I018301/1: 'The Use of Paradata in Cross-Sectional and Longitudinal Research' and Workpackage 1 of the ESRC National Centre for Research Methods (2014-2019), grant number ES/L008351/1. Vassallo's work was supported by the University of Southampton, School of Social Sciences Teaching Studentship and by the UK Economic and Social Research Council (ESRC), PhD Studentship (ES/1026258/1).

Abstract

Multilevel multiple membership models account for situations where lower level units are nested within multiple higher-level units from the same classification. Not accounting correctly for such multiple membership structures leads to biased results. The use of a multiple membership model requires selection of weights reflecting the hypothesized contribution of each level two unit and their relationship to the level one outcome. The Deviance Information Criterion (DIC) has been proposed to identify such weights. For the case of logistic regression, this study assesses, through simulation, the model identification rates of the DIC to detect the correct multiple membership weights, and the properties of model variance estimators for different weight specifications across a range of scenarios. The study is motivated by analyzing interviewer effects across waves in a longitudinal study. Interviewers can substantially influence the behavior of sample survey respondents, including their decision to participate in the survey. In the case of a longitudinal survey several interviewers may contact sample members to participate across different waves. Multilevel multiple membership models are suitable to account for the inclusion of higher-level random effects for interviewers at various waves, and to assess, for example, the relative importance of previous and current wave interviewers on current wave nonresponse. To illustrate the application, multiple membership models are applied to the UK Family and Children Survey to identify interviewer effects in a longitudinal study. The paper takes a critical view on the substantive interpretation of the model weights and provides practical guidance to statistical modelers. The main recommendation is that it is best to specify the weights in a multiple membership model by exploring different weight specifications based on the DIC, rather than prespecifying the weights.

Key words: deviance information criterion, interviewer effects, multilevel multiple membership models, survey nonresponse

Introduction

In interviewer-administered surveys interviewers can substantially influence the behavior of respondents, including their response to the survey participation request, and that is the case in both cross-sectional (Blom et al., 2010; Durrant & Steele, 2009; Durrant et al., 2010; Durrant & D'Arrigo, 2014, West and Blom, 2016) and longitudinal surveys (Campanelli & O'Muircheartaigh, 1999; Pickery & Loosveldt, 2002; Pickery et al., 2001; Haunberger, 2010; Lynn et al., 2013; Vassallo, Durrant, Smith and Goldstein, 2015; Vassallo, Durrant and Smith 2017; Brunton-Smith et al. 2016).

Interviewers influence respondents by introducing the survey concept, engaging the respondent, addressing any queries, and ultimately gaining response (Groves & Couper, 1998; Hox & De Leeuw, 2002). The resulting interviewer variability introduces non-zero correlations (or clustering) in the responses among sample units worked on by the same interviewer. These within-interviewer correlations, however, reduce effective sample sizes, similar to cluster sampling. West and Blom (2016) report that an average interviewer workload of 35 respondents and a within-interviewer correlation of only 0.03 would double the estimated variance of a mean, or what is effectively the same halve the sample size, which stresses the importance of understanding interviewer-level characteristics and other factors that introduce this type of variability in different survey outcomes. A better understanding of such interviewer influences and behaviours is therefore important in helping to reduce nonresponse in surveys before or during data collection and also for improving response propensity models. Interviewer effects may be complicated for longitudinal surveys. Across the waves of a longitudinal survey more than one interviewer may contact sample members to participate in a survey. A modeling problem particular to this kind of

data pertains the influence of *several* interviewers (i.e. the inclusion of higher-level random effects for interviewers) across various waves, whilst a substantive problem is, for example, the assessment of the relative importance of previous and current wave interviewers on current wave nonresponse (attrition) (Vassallo et al. 2015; Pickery et al, 2001; Lynn et al. 2013). If all the 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 between interviewer variance (Goldstein, 2011a) with significant biasing effects on parameter estimates in response propensity models (Chung and Beretvas, 2012).

One approach to correctly handle this data structure is to use multiple membership (MM) models (Lynn et al., 2013). Multiple membership models account for situations where lower level units are nested within multiple higher level units from the same classification. Not accounting correctly for such multiple membership structures would lead to biased results. For example, ignoring the structure and assigning each lower level unit to just one of their higher level units and then fitting the nearest equivalent hierarchical model to multiple membership data will lead to misattributed response variation to the included levels (van Landeghem et al., 2005; Moerbeek, 2004; van den Noortgate et al., 2005; Tranmer and Steel, 2001). This may lead us to draw misleading conclusions about the relative importance of different sources of influence on the response variable. Vassallo et al. (2015) compare cross-sectional and multiple membership models in accounting for different interviewers across waves using data from the Family and Children Study. Multiple membership 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 and have been used to interpret the influence of interviewers on, for example, nonresponse. The choice of weights can either be based on theory, when a strong

theoretical basis exists, or an empirical assessment using the Deviance Information Criterion (DIC), as proposed in Goldstein (2011a) and advocated in Lynn et al. (2013).

Although the multiple membership methods are in this paper applied to 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 many other behavioral sciences 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 multiple neighborhood effects on the propensity to seek traditional birth assistants in sub-Saharan Africa, where neighborhood effects have a multiple membership structure, in that both the actual neighbourhood one resides in, but also adjacent neighborhoods may influence's one's views regarding health care decisions; studies on the influence of religious group affiliation on the likelihood of undertaking volunteering work; receipt of unemployment benefits with changing household membership in longitudinal studies; and veterinary studies considering the influence of flock memberships on disease contagion. Multiple membership models have been employed in the analysis of the impact of area of residence on individual health outcomes (Chandola et al., 2005), the impact of teachers' input on student educational outcome (Fielding, 2002) and the impact of chickens' membership formations on the spread of salmonella (Rasbash & Browne, 2001).

An important question in multiple membership model applications is regarding the model specifications. The properties of parameter estimators can be sensitive to such model specifications, particularly to the omission or misspecification of the higher-level structure (Chung

& Beretvas, 2012; Luo & Kwok, 2009; Meyers & Beretvas, 2006; Tranmer & Steel, 2001). 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, the resulting properties of the estimators and the feasibility of a substantive interpretation of the chosen model need to be investigated.

The DIC is a Bayesian model selection tool which takes into consideration both the goodness-of-fit and the complexity of the model. It is particularly appropriate for models including hierarchical parameters estimated using Markov Chain Monte Carlo (MCMC) methods (Spiegelhalter et al., 2002). Some authors have analyzed the performance of the DIC for different model types and subject areas (Zhu & Carlin, 2000; Berg et al., 2004; Wilberg & Bence, 2008; Kizilkaya & Tempelman, 2003; Ward, 2008), also via means of simulation studies. These studies generally show that the DIC measure performs well in detecting the true model or similar models which adequately represent the data. However, there is very limited literature which explores the estimator properties and power of a significance test for MM models. Browne et al. (2001) investigate the properties of model estimators for MM models using a simple simulated education data example. 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 authors only consider a case for true MM weights of 0.5 and 0.5, and specify the model weights to be the correct weights. The estimator properties in the case of incorrectly specified model weights are not considered. Although literature in this area is very scarce, Wolff Smith and Beretvas (2014) investigate the choice of weights in multiple membership models with a continuous dependent variable. They compared parameter estimates and residual

estimates resulting from use of different weight patterns using a real dataset and a small-scale simulation study. Several conditions were manipulated in this study, including the mobility rate (percent of students that changed schools), intra-class correlation coefficient, number of schools and number of students per school. They found that the choice of weights does not greatly impact parameter estimates. Some studies using MM models with real data to investigate substantive questions make some reference to the robustness of the parameter estimates across different weight specifications (Fielding, 2002; Goldstein, 2011b). 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 would be expected to be biased and confidence intervals to have poor coverage properties.

This paper assesses, through simulation studies, the model identification rates of the DIC to detect the correct MM weights for a two-wave design and a binary outcome, and the properties of model variance estimators for MM models with different weight specifications across a range of scenarios. These two assessments have not been undertaken in previous literature. The correct model identification rates of the DIC are assessed in terms of the percentage of times the models with the correct weights correspond to the lowest DIC value. The properties of the variance estimator considered include the percentage relative bias, the confidence interval coverage, the power of the Wald test and the 95% credible interval estimates for the random effects parameter. The properties of the MM models are investigated when weights are chosen a priori and alternatively when chosen on the basis of the DIC. The different scenarios considered vary in

terms of the true MM weights, the different profiles of interviewer change and the proportions of cases experiencing interview change. These profile types aim to represent different plausible interviewer allocations, with the intention of covering the main possible interviewer work allocations. They reflect the various possible scenarios that induce interviewer change in surveys. Different total sample sizes and number of interviewers (groups) are also considered.

Rather than mathematical theoretical derivations, this paper uses simulation studies. Previous methodological research shows that even for perfectly nested two-level hierarchical models, any power formulae are approximate (Snijders, 2005). To derive theoretical formulae for the various properties for cross-classified and multiple membership logistic models has proved impossible. Instead, simulation studies offer a general procedure for estimating power and other estimator properties in complex designs. The only disadvantage is that a great number of simulations are required to cover a variety of possible factor values, and this is very time consuming. Other studies have also used simulation studies to investigate the impact of various factors on the properties of estimator for perfectly nested hierarchical models (Rodriguez & Goldman, 1995; Paccagnella, 2011; Moineddin et al., 2007; and Theall et al., 2010); cross-classified models (Browne & Golalizadeh, 2009) and multiple membership models (Browne et al., 2001; and Chung & Beretvas, 2012). Throughout the simulation study, the example of interviewer effects in a longitudinal survey serves as a running example.

After the simulations, an application from survey methodology is discussed, implementing multiple membership models with different weight specifications in waves 7 and 8 of the UK Family and Children Survey to investigate the relative importance of previous and current interviewers on current wave nonresponse in a longitudinal survey. Whilst the resulting practical survey design implications are not the main focus of this paper, our applied work will lead to

improved response propensity models for longitudinal surveys. Such models can then be used for either nonresponse adjustment in survey estimates (e.g. Skinner and D'Arrigo, 2011) and/or for the improvement of survey management processes, such as adaptive and responsive survey designs (Groves and Heeringa, 2006; Durrant, et al. 2015).

The implications of the work in this paper are wide ranging. The simulation results 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 highlight any inference problems arising for MM models. The performance of the DIC in choosing between competing MM model weights indicates whether the substantive interpretation of the weights based on the DIC can be emphasized. This study also indicates under which scenarios choosing weights based on an empirical assessment method compared to relying on predetermined weights yields better estimator properties and power of the Wald test. In particular, this paper provides practical guidance to users applying MM models in their work. The research may also inform the design of studies with MM structures.

Methodology

The Multiple Membership Model

Let $y_{ij_{\mathrm{p}}j_c}$ denote the dependent binary variable of interest for individual *i* nested within the two higher level multiple membership units j_{p} and j_c. For the example of survey response in longitudinal surveys, this indicates whether individual *i*, interviewed by interviewers j_{p} in the previous wave and j_c in the current wave, responded to the survey request at the *current* wave.

The probability of an individual experiencing the event of interest, here survey response, is denoted $p_{ij_pj_c} = \Pr(y_{ij_pj_c} = 1)$. Modeling this probability the logistic multiple membership multilevel model can be written as (Goldstein, 2011a):

$$\text{logit}\left(p_{ij_pj_c}\right) = \beta_0 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_{ij} + \mathrm{w}_{ij_p}u_{j_p} + \mathrm{w}_{ij_c}u_{j_c}\ , \quad \mathrm{w}_{ij_p} + \mathrm{w}_{ij_c} = 1, \qquad (1)$$

where β_0 represents the overall intercept in the linear relationship between the log-odds of *y* and the predictor variables included in the model $\mathbf{X}_{ij}$, where $\boldsymbol{j}$ represents j_pj_c. The vector $\boldsymbol{\beta}_1$ contains the parameter coefficients for each explanatory variable in the model. The interviewer-specific residuals, u_{j_p} and u_{j_c}, for both the current and previous wave interviewers, come from one distribution $N(0, \sigma_u^2)$. Cases experiencing interviewer change have a weighted average effect of the previous and current wave interviewer effects. The terms w_{ij_p} and w_{ij_c} represent the respective weights for the interviewers at the previous and current wave. 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 summing to 1. In the case of the weights w_{ij_c} being set to 0 for *all i* the multiple membership model would reduce to a simple 2-level model. It should be noted that in multiple membership multilevel models the weights are fixed quantities that need to be specified in advance and are not estimated by the model (Goldstein, 2011a). For the particular example of investigating nonresponse influences at the *current* wave it should be noted that in survey practice additional types of (unit-) nonresponse exist that occur at all previous waves (including at wave 1). In the example above, however, the nonrespondents from previous waves are not considered part of the analysis sample, i.e. attrition effects between wave 1 and wave p are not taken into account. If a researcher wanted to analyse the impact of

interviewer effects on *total* drop-out (i.e. the changes in the sample between wave 1 and the current wave), these selection effects would need to be considered.

The Simulation Study

The simulation design is as follows. First the data generating process is described. Then, the MM multilevel logistic regression model fitted to the simulated data is presented. Next, various simulation scenarios and the design factor values are considered. The formulae used to calculate the properties of the estimator, and the correct model identification rates of the DIC from these stored quantities are presented. All parameter specifications in the simulation study, where possible, are based on data from the UK Family and Children Survey, to ensure as realistic estimates and scenarios as possible in the design of the study.

Data Generating Procedure

Since the study focuses on the properties of the estimator for the interviewer random parameter only, an empty model without covariates is sufficient. The regression coefficient for the overall intercept β_0 is determined from the overall probability of the outcome variable for the mean interviewer membership, π, as:

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

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. 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 1. W_{ij}, equivalent to $W_{ij_pj_c}$, refers to the pair of true MM weights for cases experiencing interviewer change. These non-zero weights are maintained constant across all cases experiencing interviewer change. Different true weight profiles W_{ij} are considered, one giving equal weights, W_{ij}=(0.5, 0.5), and the others giving unequal weights W_{ij}=(0.9, 0.1) and W_{ij}=(0.7, 0.3). 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:

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

These values are then converted to probabilities:

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

Values of the dependent variable y_{ij} for each case, are generated from a Bernoulli distribution with probability p_{ij}. For each scenario 1000 simulations are generated using R Version 2.11.1. (It should be noted that the conditioning on prior wave response is not of relevance for the design of the simulation study.)

Simulation Model

The following model is fitted to each simulated dataset:

$$\text{logit}\left(p_{ij_pj_c}\right) = \beta_0 + w_{ij_p} u_{j_p} + w_{ij_c} u_{j_c} , \quad (5)$$

$$w_{ij_p} + w_{ij_c} = 1. \quad (6)$$

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). (The choice of weight profiles are motivated by a research question in survey methods, if the earlier or the later interviewer in a longitudinal study is more important in gaining response from sample members. Findings in the literature that range from identifying the first interviewer in the first two waves of a longitudinal study having the largest influence (Pickery et al., 2001) to identifying the later interviewer to be more important (Lynn et al. 2013)). 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 predefined weights. After all 9 models are fitted, the model 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 used to fit the models to the simulated data. Models are fitted using the MCMC estimation method with the default priors (these are diffuse/uninformative priors), a burn-in length of 5,000 and 100,000 iterations. (Different burn-in lengths were explored 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, 2017) were checked to determine how long the chain must be run for accurate point estimates and 95% credible intervals.) The second order penalized quasi-likelihood (PQL) estimates provide initial values for parameters. Multiple membership models in MLwiN require the specification of the weights by the user. For each model run the Brooks-Draper diagnostic, and the lower and upper bound of the Raftery-Lewis diagnostic were obtained and saved. For each scenario the percentage 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, indicating the percentage of times convergence was reached were calculated. (All other settings of the Brooks-Draper and the Raftery-Lewis diagnostics use the MLwiN default settings, further details are provided in the Appendix.) For all cases the percentage of times full convergence was reached was always at least 90%, which was considered acceptable. (It should be noted that convergence here is assessed within a comprehensive simulation study and a trade-off between computing time and the percentage of cases that have converged need to be set. Setting the cut-off value even higher would have resulted in significant longer computing time and consideration of less scenarios.)

Simulation Scenarios and Factors

A wide range of different scenarios, motivated by design choices and realistic survey design settings from survey practice, are considered and 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. All values were chosen based on estimated parameters using the longitudinal Family and Children Study (further details on the data see Vassallo et al. 2015) so to reflect realistic settings. (The estimated random interviewer effect of 0.3 from the data reflects a significant interviewer effect, with 8% of the total variation due to interviewers (calculated using the threshold model definition) (Goldstein, 2011a)). 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, test statistic and DIC for realistic general household survey scenarios.

Six change profile types (A-F) are considered here and their main characteristics are outlined in Table 1. These profile types aim to represent different plausible, yet extreme, interviewer allocations, with the intention of covering the main possible interviewer work allocations. They reflect the various possible scenarios that induce interviewer change in surveys. (This may range from for example no interviewer change to interviewer change occurred since a previous interviewer dropped out of the workforce and the hence the interviewer workload was reallocated to other interviewers). For all scenarios the caseload for all interviewers in the previous wave is 24 cases. 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 of cases (represented by the percentage change factor) from the previous wave. These are allocated randomly to different interviewers in the current wave. The Type B scenarios include a

new pool of interviewers at the current wave to whom cases with interviewer change are allocated randomly. 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. 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. 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 changed 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 the group of dropped interviewers.

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. Table 1 indicates the total number of caseloads per interviewer (Column titled: Current wave: pool of interviewers and case load). When interpreting the results we need to bear in mind that the level 2 units in the current wave vary amongst some profiles as outlined in table 1. (Here in the simulation we have varied case loads for a *group* of interviewers at a time rather than, what is strictly speaking the case in survey practice, allowed for different case loads for each interviewer. However, this will not affect the principle results of the simulation. It should also be noted that multilevel models naturally take account of different cluster sizes per higher grouping variable.)

[Table 1 about here]

Properties of the Variance Estimator and Correct Model Identification Rates for the DIC

The properties of the variance estimator considered include the percentage relative bias, the, the confidence interval coverage (from both the 95% Wald confidence intervals and 95% credible interval estimates), and the power of the Wald test. The correct model identification rates for the DIC are calculated as follows. For each scenario, 1000 simulated datasets are generated. For each of these 1000 datasets 9 models are fitted, each specifying different pre-defined w_{ij}. For each simulation run, out of these nine models the model corresponding to 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 investigated. The proportion of times the model with the correct model weights (w_{ij}=W_{ij}) is selected represents the correct model identification rate of the DIC measure. A less strict measure quantifies the percentage of times the

correct model weights or the adjacent model weights (e.g. for W_{ij}=(0.5, 0.5) adjacent weights would be w_{ij}=(0.4, 0.6) and w_{ij}=(0.6, 0.4), are selected.

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

$$\frac{1}{1000}\sum_{i=1}^{1000}\frac{\hat{\theta}_i-\theta}{\theta}*100\ ,$$

where $\hat{\theta}_i$ is the parameter estimate, θ is the true parameter value and i is the simulation number. The confidence interval coverage rate (see, for example, Maas & Hox, 2005) 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 nominal 95% rate. The results from the 95% credible confidence interval from the MCMC chains are provided in the appendix. 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. The proportion of datasets for which the null hypothesis is retained is subtracted from 1 to obtain the power of the test. These properties are estimated ten times – nine of which correspond to the models with prespecified w_{ij} 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}) are compared to the models with the other eight incorrect models with prespecified w_{ij} as well as the model with weights based on the DIC.

Results of the simulation study

Given the wealth of results, based on the wide range of factors and scenarios considered, this section presents key results only and outlines general trends in the properties of the estimator, the

power of the Wald test and the correct model identification rates of the DICmeasure. Full modeling results are provided in an online Appendix, with some commentary, for the interested reader. As a baseline, the properties for the model specifying the correct weights, that is $w_{ij}=W_{ij}$, are highlighted in all tables. The tables illustrate how misspecification of weights compare to the true weights within change profiles.

Percentage Relative Bias of the Level 2 Variance Estimator

The percentage relative bias of the variance estimator is evaluated for different simulation factor values. Negligible or low relative percentage bias (of less than 4%) is observed for models specifying the correct model weights w_{ij} (highlighted) 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. Table 2 shows that model weight misspecifications have greater negative consequences for the percentage relative bias of the variance estimator for scenarios with a higher proportion of interviewer change (proportion of cases with multiple memberships).

[Table 2 about here]

Table 3 shows that generally, for the W_{ij}=(0.5, 0.5) scenarios, symmetry in the distribution of the absolute values of the biases around the w_{ij}=(0.5, 0.5) model can be observed, with the lowest bias obtained for the model specifying the correct model weights [w_{ij}=(0.5, 0.5)]. 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, some skewness is observed for change profile types with unequal numbers of interviewers and unequal workloads across the two waves (Type B, C and E

scenarios). On the other hand, Table 3 shows that for W_{ij}=(0.9, 0.1) 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. The point at which the bias turns negative varies by change profile type. Although low biases are observed where the positive bias turns negative, the average DIC consistently shows higher values as the discrepancy between the w_{ij} and the W_{ij} increases (see Online Appendix).

For some scenarios with W_{ij}=(0.9, 0.1) and a low percentage of multiple memberships or a very restrictive change profile, symmetry in the biases [usually noticeable only for W_{ij}=(0.5, 0.5) since the degree of misspecification is symmetrical around w_{ij}=(0.5, 0.5)] is observed across the different models with different weights (Table 3). 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).

A lot of variation in bias across different change profile types can be observed for the models including the most incorrect w_{ij} (Table 4). However, the bias of the estimator across the different change profile types is relatively stable for the models including the correct and neighboring w_{ij}.

[Table 3 about here]

Importantly, a low relative percentage bias is obtained when the w_{ij} choice is based on the DIC. Basing the selection of the weights on the DIC avoids the possibility of huge biases in the interviewer variance if weights are considerably misspecified. Moreover, for equally distributed true MM weights [W_{ij}=(0.5, 0.5)], for all change profile types except Type F, specifying the

correct weights does not offer a major improvement in terms of the estimator bias compared to choosing the weights on the basis of the DIC (Table 3). In contrast, substantially higher biases are obtained for the models including weights based on the DIC compared to the models including the true predefined weights profile for W_{ij}=(0.9, 0.1). 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 predefined misspecified weights for the scenarios considered.

Although not the main focus of the study, we also consider the effect of different weight specifications on the bias of the regression coefficient. We find that across all scenarios considered, the relative percentage bias was throughout very small and never more than 3%, which was the case under severe misspecification of the multiple membership weights (i.e. when the weights are (0.1, 0.9) but the true set of weights are the opposite extreme (0.9, 0.1)). This indicates that the effects of setting the weights on level-one parameter estimates are very small. This is in line with results in the (scarce) literature in this area, which found that there are no substantial effects on level-one variables even if the multiple membership structure is ignored (Chung and Beretvas, 2012).

Coverage of the Confidence Interval for the Level 2 Variance

The results presented here are based on the 95% Wald confidence interval. The 95% credible confidence interval from the MCMC chains are provided in the Appendix. (The two measures show similar values and the same trends across factors.) Models specifying the correct 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%, confirming the result presented by

Browne et al. (2001) for their simulated example. Consequently, the confidence interval coverage rates for the models with the correct or neighboring 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.

There is some indication that for models with prespecified 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}. 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.

[Table 4 about here]

Table 4 shows that, as expected, W_{ij}=(0.5, 0.5) scenarios show symmetry in the confidence interval coverage around the (0.5, 0.5) weights model. Some skewness (similarly to the skewness observed for the bias of the estimator) is observed for change profile types with unequal numbers of interviewers and unequal workloads across the two waves (Types B, C and E). Interestingly, the average DIC value across the different models with different prespecified w_{ij} shows perfect symmetry for all change profiles types (see Online Appendix). In the case of W_{ij}=(0.9, 0.1) scenarios the coverage rates remain relatively high and 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 (Table 4). For the unequally distributed W_{ij} scenarios [W_{ij}=(0.9, 0.1)] the change profile types including a higher number of total interviewers (480 interviewers for Type B, and 360 interviewers for D and F) obtain better coverage rates than the change profile types with a total of 240 distinct interviewers (Type A, C and E) for the models with incorrect weights.

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 4). However, for this scenario coverage rates fall to 73% for models including incorrect prespecified 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 true multiple membership weights are unknown. Given that the simulations are based on 1000 replicates, the error rate of the estimated coverage rates is at most +/- 3 percentage points.

Power of the Wald Test for the Level 2 Variance

The power of the Wald test 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 (proportion of cases being associated with multiple memberships) and small total sample sizes N. 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. The effect of N is only noticeable for high percentage change values and different W_{ij} for different change profile type scenarios. It is important to note that the models including weights based on the DIC always obtain optimal power values (greater than 0.95).

Correct Model Identification Rates for the DIC Measure

This section explores the rates at which the DIC chooses the model with the correct multiple membership weights w_{ij} (correct model identification rates) Figure 1 shows the frequency distribution of the w_{ij} specified for the 1000 models (out of the 9000 models of each scenario) corresponding to the lowest DIC. Table 5 shows 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.

DIC performs better for scenarios with a greater percentage of cases with multiple memberships. 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 are obtained for scenarios with a greater percentage of cases experiencing interviewer change.

[Figure 1 about here]

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 to lead to the correct choice of weights w_{ij} for the W_{ij}=(0.5, 0.5) scenarios (the correct weights are always most closely associated with the lowest model DIC) . However, the effect of the sample size n on the correct model identification rates of the DIC measure varies by change profile type and by W_{ij}.

As can be observed in Table 5 the correct model identification rates of the DIC vary by true weights W_{ij}, showing better results for unequally distributed weights data, noticeable to a greater extent for scenarios with a low percentage of cases with multiple memberships. This higher DIC correct model identification rate 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. 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 5 about here]

Table 5 compares the correct model identification rates of the DIC for different change profile type scenarios with typical values for the other factors. The change profile types that do not include new interviewers in the current wave (Type A, C and 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.

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 neighboring weights are considered acceptable.

Application of multiple membership models to the analysis of interviewer effects

To demonstrate the use of multiple membership models using different weighting schemes and the DIC as an assessment criteria in a practical application setting the modelling approach is now applied to the analysis of interviewer effects on nonresponse in a longitudinal study. Interviewers have a crucial role in gaining response from sample members (Pickery et al., 2001; Durrant and Steele, 2009; Durrant et al, 2010; Haunberger, 2010). To analyse such interviewer effects a multilevel modelling approach has been advocated (Hox, 1994; O'Muircheartaigh and Campanelli, 1999). However, the analysis of interviewer effects can be greatly complicated for longitudinal studies. For example, it is unknown how interviewers affect nonresponse behavior in a longitudinal study across waves. Whilst some sample cases keep the same interviewer across time, some will experience a change in interviewers.

A key research question, which is discussed in this application, is if the interviewer from a previous wave or the current interviewer has the larger effect on nonresponse. A range of multilevel multiple membership models are applied, which explore the different weights allocated to each interviewer. One study exists, which found that the first interviewer in the first two waves of a longitudinal study has the largest influence (Pickery et al., 2001). However, this study applied a cross-classified multilevel model, a method that does not naturally lend itself to this type of application. In particular, the model makes an independence assumption and does not account for the fact that some interviewers remain the same across waves. (For a detailed comparison of the use of cross-classified and multiple membership models to longitudinal studies see Vassallo et al. 2015). The study also did not analyse interviewer effects on nonresponse at a later stage of a longitudinal study. In the application here we propose the use of a multiple membership model with varying weight specifications to disentangle the differential influence of two consecutive wave interviewers on nonresponse.

We use data from wave 7 and wave 8 of the UK Family and Children Survey, which collects information on the health and socio-economic status of households with children in the UK. An advantage of these data is that detailed information on interviewers is available and linked to the two congruent waves of the survey, which can be used to explain the interviewer effect. The information on interviewers is obtained via administrative data on interviewers recorded by the National Centre for Social Research (NatCen) and via a separate survey of interviewers also carried out by NatCen (for further information about the data see Lyon et al., 2007 and Vassallo et al. 2015). The main outcome of interest is whether or not a person responded to wave 8, conditioning on response to wave 7. This allows detailed information on both the respondents and the nonrespondents to wave 8 to be obtained from the previous wave. The final analysis sample

includes 5932 cases pertaining to 307 wave 7 interviewers, and 275 wave 8 interviewers. 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. (Further detail about the data and the survey design is described in Vassallo et al. 2015).

Application of multiple membership models with different weight specifications

The multiple membership model given in equation (1) is applied to wave 7 and 8 of the study, with the response indicator (1 for nonresponse and 0 for response at wave 8) as the dependent variable. First, we explore the multilevel random structure (as is usually done in multilevel modelling not yet including covariates (Goldstein, 2011a), i.e. we estimate an unconditional multilevel multiple membership model). The models are estimated using MCMC in MLwiN with default priors (these are diffuse/uninformative priors), a burn-in length of 5,000 and 500,000 iterations. Again, different burn-in lengths were explored to identify the appropriate length of discarded iterations to avoid undue influence from the starting values (Gelman et al., 2004).

Table 6 shows the DIC, and estimates of the random (and fixed) effects for a range of multiple membership models under the different weight specifications for wave 7 and wave 8 interviewers. The Raftery-Lewis and Brooks-Draper and diagnostics (Browne, 2017) are provided in the appendix (Table A14) (with an explanation on these diagnostics at the beginning of the Appendix). For model 7 with weight specification (0.9, 0.1) the Raftery-Lewis diagnostics for the variance is (33717; 14127) and the Brooks-Draper diagnostic is 30641, indicating convergence. Models that allocate a relatively equal weight between the two interviewers seem to perform somewhat worse. Of the multiple membership models, the smallest DIC is obtained for the model

that allocates the largest weight to the most recent interviewer (wave 8; weight specification (0.9; 0.1). Table 6 also shows the change in the DIC in comparison to this model. It should be noted that the differences in the DIC between models are relatively modest. For our data analyzing interviewer effects on nonresponse at a later stage of a longitudinal survey the analysis therefore provides some indication that the most recent interviewer seems to have the highest influence on nonresponse. We found that the wave 8 interviewer is significant and accounts for about 8% of the total variance. Exploring different multiple membership model settings we observe that different weight specifications only very slightly affect the fixed effects parameter estimates (see Table 6) and none of the changes are significant. This is in line with our earlier findings and the literature in this area that has found no substantive effects on level-one covariates even if the multiple membership structure is ignored (Chung and Beretvas, 2012).

Once an appropriate random multiple membership model structure specification is identified, it is the aim in a multilevel model to explain (part of) the significant random variance structure, here in this application the interviewer effect, by including (groups of) explanatory variables. Variables can also be included as controls. Our application makes use of unusually rich auxiliary information which allows the investigation of the influence of information on interviewer socio-demographic characteristics, work history, personality traits and job attitude. The model further controls for a range of survey design and participation history variables, as well as individual and household level characteristics. We found in this application that certain interviewer characteristics have a higher association with nonresponse. This is the case, for example, for less experienced interviewers, and or cases that experience an interviewer change after controlling for household effects. Interviewer personality traits did not explain much of the interviewer variance. A full interpretation of a similar multiple membership model investigating the influence of the various

explanatory variables from a substantive perspective has been described in the paper by Vassallo et al. (2015). We also compared parameter estimates across models with covariates under different weight specifications and found no substantive differences, as we may expect given our earlier findings and the results in the literature in this area (Chung and Beretvas, 2012).

[Table 6 about here]

Conclusions and Guidance for Modeling Practice

This paper investigates the properties of the variance estimator and the test statistic for multiple membership models when the true multiple membership weights are unknown, as would be the case in a real life situation, and how such properties change depending on the model selection method chosen. Different multiple membership models with various weight specifications are considered. The models include possible prespecified weights, and models based on the weights identified as giving the best fit by the DIC. Also, the correct model identification rates of the DIC in identifying the true multiple membership weights is examined. 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 would be expected to be biased and confidence intervals to have poor coverage properties. An application is provided, where a range of multiple membership models with different weight specifications are fitted to data from the UK Family and Children Survey to investigate the effects of interviewers on nonresponse in a longitudinal study. Research in this area is scarce. We are aware of Wolff Smith and Beretvas (2014) who analysed the continuous case using a small-scale simulation study and found that under the conditions examined in their study, choice of weight pattern did not greatly impact relative parameter bias nor level two residuals' ranks. This paper here aims to contribute to this

gap in the literature. A particular focus is on guidance to modelling practitioners, that can be derived from the results found.

The key results are as follows. As expected, the results show optimal properties for models specifying the correct model weights. The properties of the variance estimator and the test statistic generally do not vary across different factor values for the models including the correct weights. In comparison, models with misspecified weights obtain less than optimal, and at times alarmingly bad results. Specifically, the results indicate that for most scenarios, models specifying equal weights (0.5, 0.5) for data with extreme unequal true weights obtain higher biases and lower coverage rates than the correctly specified models. DIC-based weights models obtain good results overall, sometimes reaching values equivalent to the models that include the correct weights. The different factors interact with each other in a complex way influencing the properties of the estimator and test statistic, and the correct model identification rates of the DIC. We have also seen that the equal weights (0.5, 0.5) setting does not perform well in our application, where there is a higher weight allocated to the later interviewer. It is hence important to carefully consider the weight specification in a multiple membership model using a DIC-based sensitivity analysis (see also Chung and Beretvas, 2012).

No effect of halving the total sample size on the bias is noticeable for the sample sizes considered, which are typical for the type of social surveys considered here. The use of very small sample sizes for multilevel models, of say less than 200 and hence small group sizes, is generally not recommended (Maas and Hox, 2005; Moineddin et al., 2007; Paccagnella et al., 2011). Although not the main focus of this study, effects of different weight specifications on regression parameter estimates are found to be only slight. The simulation found at most a relative bias of 3%, and this only under severe misspecification of weights. In the application, effects on parameter estimates

are all negligible. This is in line with previous research that found ignoring multiple membership structures in multilevel models to lead to no substantial effects on level-one parameters (although severe effects on other level parameters and variances have been found) (Chung and Beretvas, 2012; see also Moerbeek, 2004).

Here in this study we have used typical scenarios and realistic parameter choices that are common in social survey settings and the findings will hence be applicable to similar settings. However, as in all simulation studies, the results are in general applicable to the factor values chosen and the scenarios considered. The results may not be extrapolated to very different survey design conditions with any certainty. Though some general trends can be observed, this study highlights the need for considering each particular application (with its particular data structure) on an individual basis to inform decisions on inference. It should also be noted, that our application considers the effects on the last two waves of the FACS data, but the application of multiple membership models to more general scenarios is straightforward. Whilst it is in principle possible in MLwiN to set weights for each sample unit, this does not seem practical (or even feasible) for the application of interviewer effects, since the weights need to be specified in advance. Other applications, such as school effects on pupils in different schools, may allow for such specifications if detailed information on the schools and school changes are available.

The results suggest that before deciding on the method to choose the weights the characteristics of the data should be noted. For example, for the case when the data include a low percentage of multiple memberships all nine models with the different pre-specified weights demonstrates good properties of the estimator and test statistic for all scenarios considered in this study. 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 the sample size and the interviewer change profile.

In general, however, the findings show that choosing weights based on the DIC criterion, despite possibly leading to multiple membership weights that may not closely reflect the true weights, results in good estimator properties and power of the Wald test. When the multiple membership weights cannot be predetermined on a strong theoretical basis, it may be best to always choose weights based on the DIC. This stimulation study has shown that whilst the DIC does not offer a very precise measure for choosing the exact correct model weights, for most scenarios the weights chosen are either the correct weights or acceptable neighboring weights. Consequently, 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 extent of variance apportionment between the higher-level groups (in our application groups of interviewers).

In the application considered, choosing the model with the lowest DIC value, led to the conclusion that the current wave interviewer has the largest influence on nonresponse. These findings indicate that for the later stages of a longitudinal survey the current wave interviewer seems to have the greatest impact on current wave nonresponse. They are in contrast with earlier findings by Pickery and Loosveldt (2002) who report that the first interviewer has the greatest influence. However, they investigated interviewer effects at the beginning of a longitudinal study, analysing wave 1 and 2 interviewers, and used a cross-classified multilevel model, and hence their result should be noted with caution. Furthermore, the substantive findings confirm that interviewer experience, grade and continuity are significant predictors of nonresponse, highlighting for example the

importance of retaining experienced interviewers within the agency, whereas interviewer personality traits are not important predictors of wave nonresponse. The findings from the applied work will lead to improved response propensity models for longitudinal surveys. Such models can be used for the improvement of survey management processes, such as adaptive and responsive survey designs (Groves and Heeringa, 2006; Durrant, et al. 2015), before, during and after data collection and for nonresponse adjustment in survey estimates (e.g. Skinner and D'Arrigo, 2011).

Reference List

Berg, A., Meyer, R. & Yu., J. (2004). Deviance Information Criterion for Comparing Stochastic Volatility Models. Journal of Business and Economic Statistics, 22, 107-20. (DOI: 10.1198/073500103288619430)

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. & 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. (DOI: 10.1177/1471082X0100100202)

Brunton-Smith, I., Sturgis, P. and Leckie, G. (2016) Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. Journal of the Royal Statistical Society: Series A (Statistics in Society), forthcoming.

Browne, W. J. (2017) MCMC estimation in MLwiN, version 3.00.

Campanelli, P. & O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey nonresponse. Quality & Quantity, 33, 59-76. (DOI: 10.1023/A:1004357711258)

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. (DOI: 10.1136/jech.2003.019539)

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. (DOI: 10.1111/j.2044-8317.2011.02023.x)

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. (DOI: 10.1111/j.1467-985X.2008.00565.x)

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. (DOI: 10.1093/poq/nfp098)

Durrant, G.B. & D'Arrigo, J. (2014). Doorstep interactions and interviewer effects on the process leading to cooperation or refusal. Sociological Methods and Research, 43, 490-518. (DOI: 10.1177/0049124114521148)

Durrant, G.B., Maslovskaya, O. & Smith, P.W.F. (2015) Modelling final outcome and length of call sequence to improve efficiency in interviewer call scheduling, Journal of Survey Statistics and Methodology, 3, (3), 397-424. (DOI: 10.1093/jssam/smv008)

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. (DOI: abs/10.1076/sesi.13.2.225.3435)

Goldstein, H. (2011a). 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. (DOI: 10.1111/j.1467-985X.2010.00657.x)

Groves, R.M. & Couper, M.P. (1998). Nonresponse in household interview surveys. New York: Wiley.

Haunberger, S. (2010) The effects of interviewer, respondent and area characteristics on cooperation in panel surveys: a multilevel approach, Quality and Quantity, 44, 957–969. (DOI: 10.1007/s11135-009-9248-5)

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.

Hox, J. J. (1994) Hierarchical Regression Models for Interviewer and Respondent Effects. Sociological Methods & Research, 22, 3, 300–18. (DOI: 10.1177/0049124194022003002)

Kizilkaya, K. & Tempelman, R.J. (2003). Cumulative t-link threshold models for genetic analysis of calving ease scores. Genet. Sel. Evol. 35, 489–512. (DOI: 10.1186/1297-9686-35-6-489)

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. (DOI: 10.1080/00273170902794214)

Lynn, P., Kaminska, O. & Goldstein, H. (2013). Panel attrition: How important is it to keep the same interviewer? Journal of Official Statistics, 30(3), 443-457. (DOI: 10.2478/jos-2014-0028)

Maas, C.J.M. & Hox, J.J. (2005). Sufficient sample sizes for multilevel modelling. Methodology: European Journal of Research Methods for the Behavioral and Social Science, 1, 85-91. (DOI: 10.1027/1614-2241.1.3.86)

Meyers, J. L. & Beretvas, S. N. (2006). The impact of inappropriate modeling of crossclassified data structures. Multivariate Behavioral Research, 41(4), 473–497. (DOI: 10.1207/s15327906mbr4104_3)

Moerbeek, M. (2004). The Consequence of ignoring a level of nesting in multilevel analysis. Multivariate Behavioral Research, 39, 129-149. (DOI: 10.1207/s15327906mbr3901_5)

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. (DOI: 10.1186/1471-2288-7-34)

O'Muircheartaigh, C. and Campanelli, P. (1999) A Multilevel Exploration of the Role of Interviewers in Survey Nonresponse, Journal of the Royal Statistical Society, Series A, 162, 3, 437–46. (DOI: 10.1177/1525822X10387770)

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. Methodology, 7(3), 111-120. (DOI: 10.1027/1614-2241/a000029)

Pickery, J. & Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. Quality & Quantity, 36, 427-437. (DOI: 10.1023/A:1020905911108)

Pickery, J., Loosveldt, G. and Carton, A. (2001) The effects of interviewer and respondent characteristics on response behavior in panel surveys—a multilevel approach, Sociological Methods and Research, 29, 509–523. (DOI: 10.1177/0049124101029004004)

Rasbash, J. & Browne, W. J. (2001). Non-hierarchical multilevel models. In Leyland, A. and Goldstein, H. (Eds.) Multilevel modelling of health statistics. Chichester: Wiley.

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. (DOI: 10.2307/2983404)

Skinner, C.J. and D'Arrigo, J. (2011) Inverse probability weighting for clustered nonresponse. Biometrika, 98, 4, 953-966. (DOI: 10.1093/biomet/asr058).

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.). Encyclopedia of statistics in behavioural sciences, (Volume 3, pp. 1570-1573). New York: Wiley.

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. (DOI: 10.1111/1467-9868.00353)

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. (DOI: 10.1136/jech.2009.097956)

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. (DOI: 10.1068/a3317)

van den Noortgate, W., Opdenakker, M.-C. & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 16, 281-303. (DOI: 10.1080/09243450500114850)

van Landeghem, G., De Fraine, B. & van Damme, J. (2005). The Consequence of ignoring a level of nesting in multilevel analysis: A comment. Multivariate Behavioral Research, 40, 423-434. (DOI: 10.1207/s15327906mbr4004_2)

Vassallo, R., Durrant, G. B., Smith, P. W.F. and Goldstein, H. (2015) Interviewer effects on nonresponse propensity in longitudinal surveys: a multilevel modeling approach. Journal of the Royal Statistical Society Series A, 178, 1, 83-99.

Vassallo, R., Durrant, G.B. and Smith, P. (2017) Separating Interviewer and Area Effects Using a Cross-Classified Multilevel Logistic Model: Implications for Survey Designs, Journal of the Royal Statistical Society, Series A, 180, 2, 531-550.

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecological Modelling, 211, 1–10. (DOI: 10.1016/j.ecolmodel.2007.10.030)

West, B. and Blom, A. (2017) Explaining Interviewer Effects: A Research Synthesis, Journal of Survey Statistics and Methodology, 5(2), 175-211. (DOI: 10.1093/jssam/smw024)

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. (DOI: 10.1016/j.fishres.2008.04.010)

Wolff Smith, L.J. and Beretvas, N. (2014) The Impact of Using Incorrect Weights With the Multiple Membership Random Effects Model, Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10, 31-42. (DOI: 10.1027/1614-2241/a000066)

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. (DOI: 10.1002/1097-0258(20000915/30)19:17/18<2265::AID-SIM568>3.0.CO;2-6)

Table 1: Change profile type characteristics covering various possible interviewer work allocations (illustrated for the case n=5760; for all scenarios 240 interviewers with 24 cases per interviewer were used in the previous wave; for type A and B the percentage change indicates the proportion of cases of each previous wave interviewer which are allocated to a different interviewer in the current wave; for 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).

Type	Short Description	Current wave: pool of interviewers and case load	Percentage change considered
A	same pool of interviewers for both waves but each interviewer loses a specific amount of cases (% change) from previous wave	240 previous interviewers (same pool of interviewers as previous wave), (total: 240 interviewers) 24 cases per interviewer	New allocation: 8%, 21%, 33%, 50%, 92%
B	new pool of interviewers at current wave to whom cases with interviewer change are allocated randomly	240 previous interviewers and 240 new interviewers, (total: 480 interviewers) 12 cases per interviewer	New allocation: 8%, 50%
C	interviewer drop-out and cases of interviewers who drop out at previous wave are distributed randomly among all other interviewers present in the previous wave	120 previous interviewers, (total: 120 interviewers) 48 cases per interviewer	Drop-out and new allocation: 50%
D	interviewer drop-out and newly recruited interviewers are allocated the changed cases randomly in the current wave	120 previous interviewers and 120 new interviewers, (total 240 interviewers) 24 cases per interviewer	Drop-out and new allocation: 50%
E	interviewer drop-out and intact caseload of a dropped interviewer is allocated randomly to another (existing) interviewer	120 previous interviewers, (total: 120 interviewers) 48 cases per interviewer	Drop-out and new allocation: 50%
F	interviewer drop-out and intact caseload of a dropped interviewer is allocated randomly to a new interviewer	120 previous interviewers and 120 new interviewers, (total: 240 interviewers) 24 cases per interviewer	Drop-out and new allocation: 50%

Figure 1: Frequency Distribution of the Model Weights for the DIC-based Weights Models (for Type A, with sample size n=5760, number of interviewers in the previous wave n_p^I=240, interviewer variance σ_u^2=0.3, overall probability of the outcome variable π=0.8, true weight W_{ij}= (0.5, 0.5) scenarios with varying percentage change in the proportion of cases of each previous wave interviewer which are allocated to a different interviewer in the current wave).

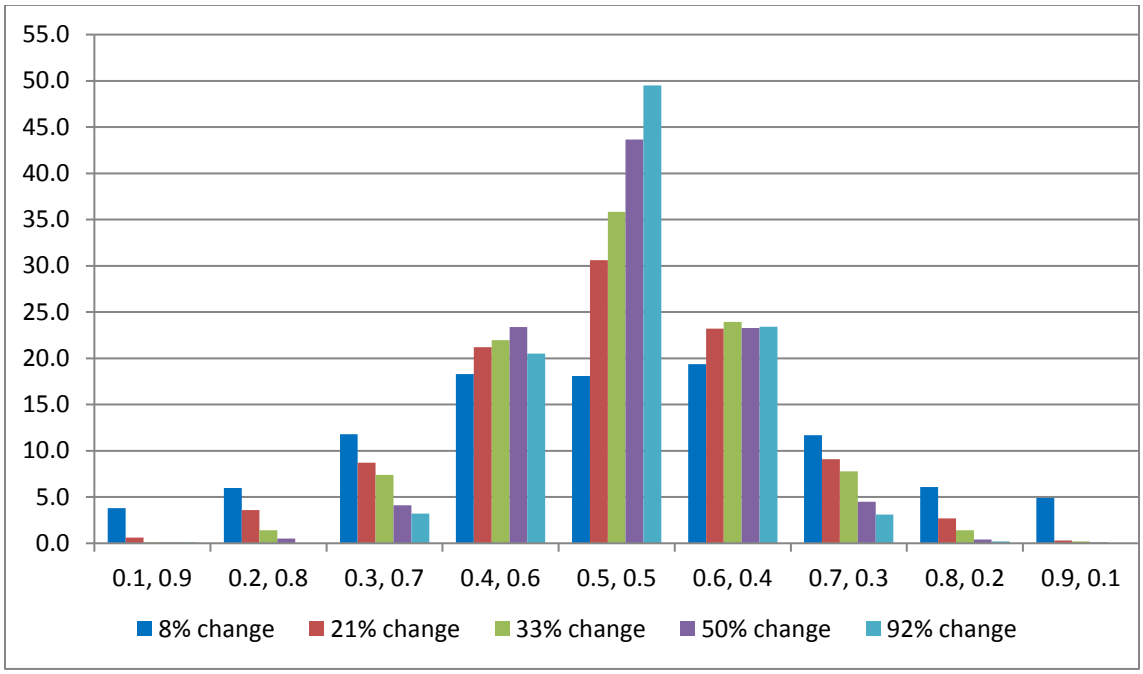

Table 2: Relative Percentage Bias (for Type A, with sample size n=5760, number of interviewers in the previous wave n_p^I =240, interviewer variance σ_u^2=0.3, overall probability of the outcome variable π=0.8, true weight w_{ij}= (0.5, 0.5) scenarios with varying percentage change in the proportion of cases of each previous wave interviewer which are allocated to a different interviewer in the current wave)

	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

Table 3: Relative Percentage Bias (for sample size n=5760, number of interviewers in the previous wave n_p^I=240, 50% change, interviewer variance σ_u^2=0.3, overall probability of the outcome variable π=0.8 with varying true weights 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

Table 4: CI coverage (for sample size n=5760, number of interviewers in the previous wave n_p^I=240, 50% change, interviewer variance σ_u^2=0.3, overall probability of the outcome variable π=0.8 with varying true weights 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

Table 5: Correct Model Identification Rates of the DIC (showing the proportion of the 1000 models that have the correct weights and the proportion which have the correct or adjacent weights for different scenarios) (for sample size n=5760, number of interviewers in previous wave n=240, 50% change, interviewer variance σ_u^2=0.3, overall probability of the outcome variable π=0.8 Scenarios with Varying Profile Change Type and true weight W_{ij})

W_{ij}	Change Profile Type	Proportion with Correct Weights	Proportion with Correct or 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 6: DIC and estimates of random and fixed effects for various multiple membership models analyzing wave 7 and wave 8 interviewer effects on nonresponse using different weight specifications (The models are ordered according to the size of the DIC).

Model	Type of Model	Wave 8 Weights	Wave7 Weights	Coefficient (S.E.)	Variance (S. E.)§	DIC	DIC Change*
1	MM	0.4	0.6	-2.411 (0.055)	0.278 (0.086)**	4159.08	5.96
2	MM	0.5	0.5	-2.410 (0.055)	0.287 (0.087)**	4159.03	5.91
3	MM	0.6	0.4	-2.411 (0.056)	0.288 (0.090)**	4158.75	5.63
4	MM	0.3	0.7	-2.415 (0.055)	0.272 (0.082)**	4158.33	5.21
5	MM	0.7	0.3	-2.414 (0.056)	0.291 (0.090)**	4157.41	4.29
6	MM	0.8	0.2	-2.419 (0.056)	0.288 (0.085)**	4155.36	2.24
7	MM	0.9	0.1	-2.426 (0.057)	0.282 (0.081)**	4153.12	-

* records the DIC change in comparison to Model 7.
§ the asterix ** refers to the Wald test.