ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model

Ian Brunton-Smith,

University of Surrey, Guildford, UK

Patrick Sturgis

University of Southampton, UK

and George Leckie

University of Bristol, UK

[Received May 2015. Final revision March 2016]

Summary. We propose a cross-classified mixed effects location–scale model for the analysis of interviewer effects in survey data. The model extends the standard two-way cross-classified random-intercept model (respondents nested in interviewers crossed with areas) by specifying the residual variance to be a function of covariates and an additional interviewer random effect. This extension provides a way to study interviewers' effects on not just the 'location' (mean) of respondents' responses, but additionally on their 'scale' (variability). It therefore allows researchers to address new questions such as 'Do interviewers influence the variability of their respondents' responses in addition to their average, and if so why?'. In doing so, the model facilitates a more complete and flexible assessment of the factors that are associated with interviewer error. We illustrate this model by using data from wave 3 of the UK Household Longitudinal Survey, which we link to a range of interviewer characteristics measured in an independent survey of interviewers. By identifying both interviewer characteristics in general, but also specific interviewers who are associated with unusually high or low or homogeneous or heterogeneous responses, the model provides a way to inform improvements to survey quality.

Keywords: Interviewer effect; Measurement error; Mixed effects location–scale model; Stat-JR software; Understanding society

1. Introduction

This paper is concerned with improving our understanding of the effects that interviewers have on survey responses in face-to-face surveys that serve to inflate the variance of parameter estimates. Interviewer behaviour can induce this effect in at least two ways: by producing differential sample compositions via their effect on response propensities (West *et al.*, 2013; West and Olson, 2010) and by influencing the answers that respondents provide during the interview (Schaeffer *et al.*, 2010). It is this latter source of interviewer error that is the primary focus of the current study. This so-called 'interviewer effect' arises through idiosyncrasies in the ways that interviewers administer questionnaires. For instance, an interviewer may repeatedly leave out the same

Address for correspondence: Patrick Sturgis, Department of Social Statistics, Murray Building, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: P.Sturgis@soton.ac.uk

word when reading a particular question or may 'help' respondents to understand an ambiguous phrase, whereas other interviewers do not (Cannell *et al.*, 1981; Kish, 1962; Mangione *et al.*, 1992; O'Muircheartaigh, 1976). In less direct ways interviewers can also influence the answers that respondents give. Female respondents, for example, may feel more pressure to give a socially desirable answer to a male than to a female interviewer, and younger respondents may answer some questions differently in the presence of an older interviewer compared with someone who is closer to their own age. Thus, interviewers may affect the responses that they obtain, not through any overt behaviour, but merely as a function of their observable characteristics (Davis and Scott, 1995).

Together, these behavioural interactions between respondents and interviewers induce a dependence in responses within interviewers which is typically expressed as an intraclass correlation coefficient (ICC). Positive ICCs increase the standard errors of parameter estimators in the same manner as multistage sampling, namely as a result of within-cluster homogeneity on survey outcomes (Hansen *et al.*, 1961; Kish, 1962). The increase in parameter estimator variance due to interviewers is typically expressed as the design effect:

$$D_{\text{eff}} = 1 + (m - 1)\rho, \tag{1}$$

where ρ is the ICC due to interviewers and m is the average number of respondents interviewed by each interviewer.

The design effect increases with the number of respondents per interviewer and, when this is large, the design effect can be sizable, even for small values of ρ. O'Muircheartaigh and Campanelli (1998), for example, found design effects as high as 5 for some items in the British Household Panel Survey, which represents a very substantial loss of efficiency. Furthermore, Schnell and Kreuter (2005) demonstrate that the interviewer component of the design effect is typically larger than the component due to area clustering. It is clearly important, then, that we understand how interviewer effects come about so that they can be mitigated through survey design, interviewer recruitment and training.

To date, interviewer effects on survey responses have almost always been conceptualized and analysed in terms of mean differences in respondents' answers with some interviewers effectively raising their respondents' 'true' answers and other interviewers lowering them. For example, recent empirical investigations of interviewer effects have fitted two-level (respondents nested in interviewers) mixed effect models (also known as multilevel models; Goldstein (2011)) to survey responses, where an interviewer random effect is included to allow the mean of the survey response, adjusted for respondent, area and interviewer covariates, to vary over interviewers, thus capturing and estimating the residual within-interviewer dependence or ICC, ρ (Hox, 1994; O'Muircheartaigh and Campanelli, 1998; West and Olson, 2010; West *et al.*, 2013). In principle, unbiased estimation of ρ requires random allocation of respondents to interviewers: a procedure that is rarely implemented in practice in face-to-face surveys for logistical and cost reasons (for exceptions see O'Muircheartaigh and Campanelli (1998) and Schnell and Kreuter (2005)). As a result, much of the existing evidence base is drawn largely from the context of telephone surveys, where interpenetrating designs are feasible. More recently, however, researchers have tended to estimate interviewer ρ by using cross-classified mixed effects models with random effects specified for interviewers and areas and which include interviewer, area and respondent level controls to adjust for non-random allocation of respondents to interviewers (Durrant *et al.*, 2010; Turner *et al.*, 2014). As with any procedure which relies on statistical control, this approach cannot guarantee unbiased estimates but comparisons between estimates by using this approach and those from randomized designs show similar patterns of effects (Brunton-Smith *et al.*, 2012).

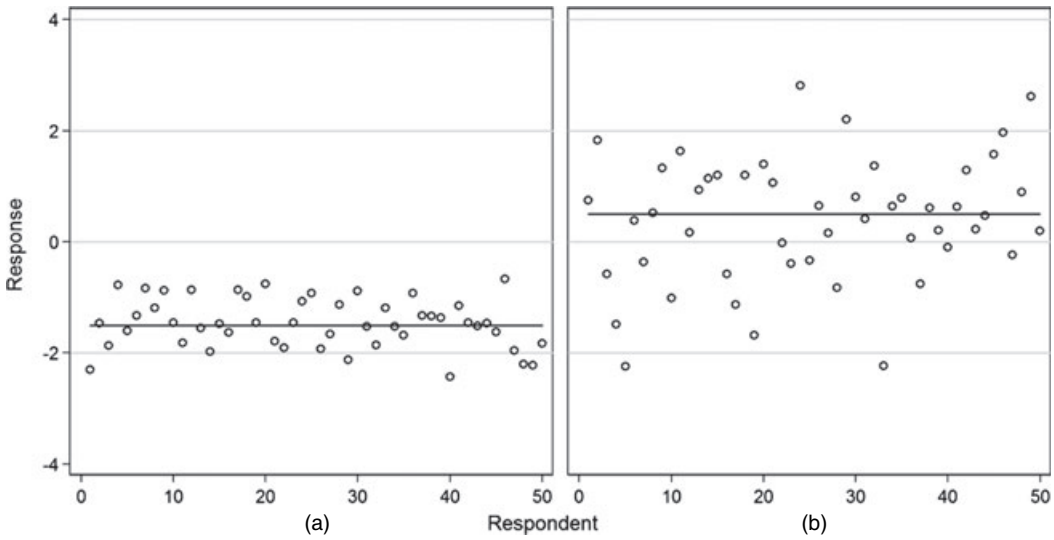

Fig. 1. Graphical illustration of respondents' answers to a hypothetical survey question for two interviewers: (a) interviewer 1; (b) interviewer 2

In addition to any effect that interviewers may have on the mean of answers that they elicit from respondents, it is plausible that they might also have an effect on the variability of respondents' answers, with some interviewers effectively amplifying the 'true' differences between respondents' answers and other interviewers dampening them. Yet existing studies, and the standard mixed effects model more generally, specify a homoscedastic residual variance and so implicitly assume that the variance of the survey outcome, having adjusted for the covariates, is constant across interviewers.

Fig. 1 graphically illustrates the difference between these two types of interviewer effect by plotting the responses (in this case z-scores) to a hypothetical survey question for 100 respondents randomly assigned to two interviewers. The horizontal lines denote the mean response for each interviewer. Interviewer 1's respondents give, on average, lower and less variable responses than those given to interviewer 2. A traditional mixed effects analysis would capture the mean differences but would ignore the differences in the variance. However, differences in variance, to the extent they might arise, clearly represent another important form of error that interviewers can introduce to survey data.

Hedeker *et al.* (2008) proposed the 'mixed effects location–scale model' to relax the homoscedastic residual variance assumption of the mixed effect model. Specifically, the standard two-level random-intercept model is extended by specifying the level 1 residual variance to be a log-linear function of the covariates and an additional level 2 random effect. Although this model was proposed for analysing intensive longitudinal data, it can equally be applied in cross-sectional settings (Leckie *et al.*, 2014), including the current case of respondents (level 1) nested in interviewers (level 2).

In this paper, we propose a cross-classified version of the mixed effects location–scale model for the analysis of interviewer effects in survey data. The model includes two interviewer random effects, to capture interviewers' potentially correlated influences on the 'location' (mean) and 'scale' (variability) of respondents' answers. An area random effect is included on the mean response to separate the influence of interviewers from the areas to which they are assigned (Brunton-Smith *et al.*, 2012; Durrant *et al.*, 2010). The model adjusts for respondent, interviewer

and area characteristics and therefore allows the analyst to address new questions such as 'Do interviewers influence the variability in addition to the average of their respondents' answers, and if so why?'. We contend that this approach provides a more complete and flexible assessment of the factors that are associated with interviewer error than existing methods. We illustrate this model by using data from wave 3 of the UK Household Longitudinal Survey (UKHLS), which we link to a range of interviewer characteristics measured in a separate survey of interviewers. We demonstrate how the model can be used to improve survey quality by identifying interviewer characteristics that are associated with more variable survey responses. We also show how this approach enables estimation of interviewer-specific ICCs, which can be used to identify interviewers with unusually homogeneous or heterogeneous responses.

2. Factors associated with interviewer effects

In trying to understand the causes of interviewer variance, existing research has focused on two primary questions: first, how different types of questions may be more or less prone to interviewer effects and, second, which interviewer characteristics are associated with larger variance components (Schaeffer *et al.*, 2010). Davis and Scott (1995) found that interviewer variance in an Australian medical survey was largest for attitudinal questions and smallest for sociodemographic variables: a pattern which has also been found using British data (Brunton-Smith *et al.*, 2012). Questions which require more input from interviewers, such as those which require the use of show cards, explanatory preambles and probing, are also subject to larger interviewer variance (O'Muircheartaigh and Campanelli, 1998; Brunton-Smith *et al.*, 2012; Mangione *et al.*, 1992). Similarly, Schnell and Kreuter (2005) found that sensitive questions, non-factual questions and open questions which require the interviewer to record 'verbatim' answers had systematically larger interviewer effects than other types of question (see also Sturgis and Luff (2015) and Collins (1980)).

Research into interviewer characteristics which drive these interviewer differences has focused primarily on easily observable demographic variables such as gender, age and ethnicity (Hox, 1994; Pickery *et al.*, 2001; Schaeffer, 1980), not least as these are often the only variables that are available on administrative databases held by survey agencies. These studies have found that although demographic characteristics do appear to be predictive of interviewer differences, the patterns of association differ quite markedly across surveys and question types. For instance, O'Muircheartaigh and Campanelli (1998) found interviewer age and gender to be significant predictors of interviewer differences for some survey outcomes in the British Household Panel Survey but not in others. Likewise, Davis and Scott (1995) found significantly larger interviewer effects among older interviewers and among those from ethnic minority groups for many but not all the items considered (see also Finkel *et al.* (1991) and Hox *et al.* (1991)). Researchers have also shown that these effects may depend on characteristics of the respondent, suggesting an interviewer matching effect (Anderson *et al.*, 1988; Kane and Macaulay, 1993; Huddy *et al.*, 1997).

In addition to these kinds of demographic characteristics, researchers have considered variables relating to interviewing experience and work performance. Using the British Crime Survey, Brunton-Smith *et al.* (2012) found that interviewers with the worst historical response rates had, on average, the largest variance components across 36 survey outcomes. O'Muircheartaigh and Campanelli (1998) found that interviewer experience and working in a supervisory capacity were significantly associated with interviewer effects (see also Bailar *et al.* (1997), Hughes *et al.* (2002) and van Tilburg (1998)). Most recently, Turner *et al.* (2014) assessed the effect of interviewer personality on outcome variance. Their rationale was that particular personality types might

be more or less prone to the sorts of behaviour that are thought to give rise to systematic differences in response variability. For example, interviewers who are higher on the conscientiousness dimension of the 'big five' personality inventory (Goldberg, 1990) may be more likely to obey instructions to read the questions exactly as they are written. Alternatively, interviewers who are high on the agreeableness, openness and extraversion dimensions may be more likely to adopt a 'chatty' and informal approach to administering the questionnaire which could, in turn, give rise to more variable responses. However, they found little or no evidence of an association between interviewer personality and response variance across a range of items in the UK National Travel Survey.

In this paper, we focus our attention on interviewer rather than question characteristics as predictors of response variance. We employ measures of interviewer demographic characteristics, survey experience and personality as predictors in our models. Additionally, we consider variables which tap interviewers' attitudes towards the value of surveys. This is based on the expectation that interviewers who place higher value on the scientific merit and practical utility of survey research will be more likely to follow the procedures and guidance that they are given about how they should undertake interviews. Where existing studies have focused only on interviewer variance inflation which is brought about via their influence on the mean of respondents' answers, we additionally consider the interviewers' influence on the variance of survey outcomes, on top of any effect that they have on the mean.

3. Analytical approach

Early methods for detecting and understanding the causes of interviewer effects used analysis-of-variance models (Bailar *et al.*, 1977; Biemer and Stokes, 1985; Fellegi, 1964, 1974). The analysis-of-variance framework is limited in its ability to estimate the effect of interviewer level characteristics on the survey outcomes accurately and to account for non-random allocation of respondents to interviewers adequately (Hox, 1994). More recently, practice has shifted to the use of mixed effects models, where a random effect is specified at the interviewer level (Pickery *et al.*, 2001; Schnell and Kreuter, 2005; O'Muirheartaigh and Campanelli, 1998; West and Elliott, 2014; West and Olson, 2010). Implementations of the mixed effects model for studying interviewer variance have also used a cross-classified extension to identify the influence of interviewers and areas separately (O'Muircheartaigh and Campanelli, 1998; Durrant *et al.*, 2010; Brunton-Smith *et al.*, 2012; Turner *et al.*, 2014).

This model has the following form. Let $y_{i(jk)}$ denote the continuous response measurement for respondent i ($i = 1, \ldots, N$) interviewed by interviewer j ($j = 1, \ldots, J$) living in area k ($k = 1, \ldots, K$), where we indicate the cross-classification of interviewers and areas by placing their indices in parentheses. The standard two-way cross-classified random-intercept model for $y_{i(jk)}$ can then be written as

$$y_{i(jk)} = \mathbf{x}'_{i(jk)}\boldsymbol{\beta} + u_j + v_k + e_{i(jk)}, \tag{2}$$

where $\mathbf{x}_{i(jk)}$ is a vector of respondent, interviewer and area level covariates with coefficients $\boldsymbol{\beta}$ and u_j and v_k are random-intercept effects representing remaining unobserved interviewer and area influences on $y_{i(jk)}$. The respondent-specific residual is $e_{i(jk)}$. The random effects and residuals are assumed to be mutually independent, independent of the covariates and normally distributed with zero means and constant variances: $u_j \sim N(0, \sigma_u^2)$, $v_k \sim N(0, \sigma_v^2)$ and $e_{i(jk)} \sim N(0, \sigma_e^2)$. The random-effect variances σ_u^2 and σ_v^2 capture the variability in adjusted mean responses across interviewers and areas respectively, whereas the residual variance σ_e^2 measures the variability in respondents' answers that is unexplained by the fixed and random effects. The ICC for

interviewers can be derived as $\rho_u = \sigma_u^2(\sigma_u^2 + \sigma_v^2 + \sigma_e^2)^{-1}$, which is the expected correlation between the responses of two independent respondents (i.e. two respondents living in two different areas) interviewed by a common interviewer.

Equation (2) assumes constant residual variance (homoscedasticity), which is to say that σ_e^2 is constrained to be constant across all interviewers and all areas. We can relax this assumption by specifying an auxiliary log-linear equation for the residual variance as a function of covariates and additional interviewer and area random effects (Hedeker *et al.*, 2008). However, given our interests here, we specify an additional random effect for interviewers only. In conceptual terms, relaxing the homoscedasticity assumption allows interviewers to influence not only the mean of $y_{i(jk)}$ but also the residual variability once any direct effects on the mean have been accounted for. The log-link function ensures that the residual variance takes positive values. This can be written as

$$\ln(\sigma^2_{e_{i(jk)}}) = \mathbf{w}'_{i(jk)}\boldsymbol{\alpha} + u_j^{[2]}, \tag{3}$$

where $\ln(\sigma^2_{e_{i(jk)}})$ denotes the logarithm of the now heterogeneous residual variance, $\mathbf{w}_{i(jk)}$ is a vector of respondent, interviewer and area level covariates with coefficients $\boldsymbol{\alpha}$ and $u_j^{[2]}$ is the additional interviewer random effect. We use the '[2]' superscript to distinguish this random effect from the usual response equation interviewer random effect in equation (2) which we now denote $u_j^{[1]}$. The two sets of interviewer random effects are assumed to be bivariate normal with zero mean vector and constant variance–covariance matrix

$$\begin{pmatrix} u_j^{[1]} \\ u_j^{[2]} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u[1]} & \\ \sigma_{u[1]u[2]} & \sigma^2_{u[2]} \end{pmatrix} \right\}. \tag{4}$$

The variance–covariance matrix summarizes the extent to which interviewers differ in both the (adjusted) mean of the answers of the respondents whom they interview (summarized by $\sigma^2_{u[1]}$) and in the variability of these answers (summarized by $\sigma^2_{u[2]}$). The matrix also captures the covariance between these two forms of interviewer influence ($\sigma_{u[1]u[2]}$).

The population-averaged residual variance, conditional on the covariates $\mathbf{w}_{i(jk)}$, is given by

$$E(\sigma^2_{e_{i(jk)}}|\mathbf{w}_{i(jk)}) = \exp(\mathbf{w}'_{i(jk)}\boldsymbol{\alpha} + 0.5\sigma^2_{u[2]}) \tag{5}$$

which can be substituted in the expression for the ICC to give the population-averaged ICC. In addition to the population-averaged ICC, it is straightforward to calculate interviewer-specific ICCs and, thereby, to identify interviewers who induce more similar responses from their respondents than other interviewers:

$$\frac{\sigma^2_{u[1]}}{\sigma^2_{u[1]} + \sigma_v^2 + \exp(\mathbf{w}'_{i(jk)}\boldsymbol{\alpha} + u_j^{[2]})}. \tag{6}$$

The model provides a flexible means of assessing the factors that are associated with interviewer-induced response variability. A notable benefit is that interviewers can have differential effects on the 'location' (the mean) and the 'scale' (the variance) of a survey outcome. So, for example, an interviewer characteristic may have a positive $\boldsymbol{\beta}$-coefficient in equation (2) and a negative or non-significant $\boldsymbol{\alpha}$-coefficient in equation (3) (or vice versa).

4. Data and measures

Data are taken from wave 3 of the UKHLS general population sample with fieldwork undertaken during 2011 and 2012. The UKHLS is a nationally representative household panel survey

comprising approximately 40 000 households at the first wave. The survey has a multistage clustered design, with a sample of postcode sectors (stratified by region, population density and minority ethnic density) selected with probability proportional to size, and 18 households then selected from each sector for interview. All residents of each selected household were eligible for interview with an average of 1.6 adults interviewed in each participating household. We use data from wave 3 because this was collected closest in time to the 'Understanding society interviewer survey'. At wave 3 a total of 30 685 full interviews were conducted with a cross-sectional response rate of 61% (Knies, 2014). Over the duration of the 24-month fieldwork period, interviewers could be assigned to multiple postcode sectors, with 668 interviewers in the field and an average of 46 interviews undertaken per interviewer.

Information about the characteristics of interviewers working on the UKHLS come from the 'Understanding society interviewer survey'. This is an on-line survey (postal for those no longer working for the data collection agency, the National Centre) of interviewer attitudes and behaviour which was fielded in the spring of 2014. Invitations were sent to all interviewers who worked on the first wave of the UKHLS ($n = 823$) and interview data were successfully obtained from 473 of them: a response rate of 58% (Burton *et al.*, 2014). The interviewer data were linked to the main UKHLS data set at wave 3. Linkage was successful for a total of 303 interviewers, who together were responsible for 17 471 interviews. In addition to age and sex, we use three questions on interviewing experience (whether interviewers had experience of working for another survey agency, non-survey interviewing or working in public engagement), three questions on beliefs about surveys ('Participation in surveys is a matter of self-interest (agree/disagree)', 'Most surveys are carried out in a responsible way (agree/disagree)' and 'In most cases survey results are correct (agree/disagree)'), and shortened versions of the 'big five' personality inventory (agreeableness, conscientiousness, extravert, neuroticism and openness). Interviewer personality traits were themselves derived from a battery of 15 survey items (see Jäckle *et al.* (2013)).

To account for the clustered sample design we use the middle layer super-output area geography (Martin, 2001). Middle layer super-output areas are preferable to postcode sectors because they are more consistent in size (containing an average of 5000 households), were constructed to maximize internal homogeneity (based on social structure) and aim to respect 'natural' physical boundaries in boundary definitions. This makes them a more meaningful spatial unit to reflect 'area' differences than postcode sectors. Middle layer super-output areas can also be easily linked to aggregate census data, enabling us to control for additional features of the local area in our models.

To illustrate the utility of the mixed effects location–scale model for estimating interviewer effects, we use three attitude questions from wave 3 of the UKHLS as dependent variables in our models. Attitudinal items were selected because previous research has indicated that they are most susceptible to interviewer influences on the location of responses (Schnell and Kreuter, 2005). The response scales for the three questions are a five-point Likert item, Q1, an 11-point scale with a more continuous distribution, Q2, and a five-point Likert scale item from the (paper) self-completion component of the UKHLS, Q3. The item from the self-completion questionnaire was selected as a way of checking that the model produces sensible results. Specifically, the model should show little or no interviewer effects because the interviewer should have little, if any, involvement in the completion of this question. The response rate to the self-completion questionnaire was 90% at wave 3 (Scott and Jessop, 2013). The question wordings for each item are as follows.

'1. *People in this neighbourhood generally don't get along with each other* (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).

'2. *On a scale from 0 to 10, where 0 means very unlikely and 10 means very likely, how likely is it that your vote will make a difference in terms of which party wins the election in this constituency at the next general election?*
'3. *The friendships and associations I have with other people in my neighbourhood mean a lot to me* (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).'

5. Estimation

We fit three models of increasing complexity for each of the three items. Model 1 is a simplified version of equation (2), including only an intercept, which we allow to vary across respondents, interviewers and areas. The response variance is decomposed into components in the usual way, except that we allow the magnitude of the residual variance to vary across interviewers through the inclusion of an interviewer random effect in the scale equation. Model 2 adds respondent and area level covariates to the location equation to adjust for uneven sample composition across interviewer assignments, which can arise because of spatial auto-correlation and differential non-response. However, since respondent level covariates will also be subject to interviewer-induced measurement errors we include only respondent gender and age. At the area level we include the following covariates: ethnic diversity, socio-economic disadvantage, urbanicity, population mobility, age and housing structure. Ethnic diversity was calculated by using the Herfindahl concentration formula (Hirschman, 1964); all other area level variables were derived by principal components analysis of aggregate census variables (see Brunton-Smith and Sturgis (2011) for details of the derivation).

Model 3 introduces the interviewer covariates. All interviewer characteristic variables are included in both the response model to capture mean differences in the outcome across interviewers and also in the residual variance model (equation (3)) to explore how response variability differs across interviewers. We allow the magnitude of the within-interviewer variance to depend on respondent gender and age. This adjusts the estimated differences across interviewers for the effects of potential respondent level heterogeneity of variance. The inclusion of a larger set of individual variables did not lead to any substantial changes to parameter estimates.

Models are fitted by using Markov chain Monte Carlo methods implemented in the Stat-JR software package (Charlton *et al.*, 2013). An explanation of how to set up the model in Stat-JR can be found in the on-line appendix. We specify diffuse (vague, flat or minimally informative) prior distributions for all parameters. All models are specified by using three chains with dispersed starting values, each with a burn-in period of 5000 iterations and a monitoring period of 10000 iterations. Visual assessments of the parameter chains and standard Markov chain Monte Carlo convergence diagnostics suggest that the length of these periods is sufficient. *QQ*-plots of model residuals confirm that normality assumptions are met, with the exception of one interviewer whose response profile is markedly different from all others when considering Q2. Data from this interviewer were omitted from analyses of Q2, although the substantive conclusions are unchanged in either case.

The UKHLS includes survey weights to correct for unequal selection probabilities when multiple households are present at each address and to adjust for attrition across waves. Currently there is no way to implement survey weights by using Markov chain Monte Carlo sampling and efforts to establish best practice are on going (Gelman, 2007). Following recommendations in Rao *et al.* (2013) we conducted a simple sensitivity analysis of our results by including the survey weight as a covariate in the model. Respondent level variables that were used in the derivation of the weight were then added as covariates and the coefficient of the weight became non-significant. This model specification did not result in any material changes to our key parameter estimates (these additional models are available on request).

We report the posterior means, standard deviations and 95% credible intervals of the 30 000 pooled monitoring iterations. These quantities are analogous to the parameter estimates, standard errors and confidence intervals from a frequentist analysis. We use the deviance information criterion DIC to compare the fit of alternative models (Spiegelhalter *et al.*, 2002); models with smaller DIC-values are preferred to those with larger values, with differences of 5 or more considered substantial (Lunn *et al.*, 2012).

6. Results

Table 1 presents the model 1 results for variables Q1 and Q2, which are taken from the face-to-face element of the survey. The model estimates a population-averaged interviewer ICC of 0.041 for Q1 and 0.028 for Q2, which are of the same approximate magnitude as ICC estimates found in comparable existing studies (O'Muircheartaigh and Campanelli, 1998; Brunton-Smith *et al.*, 2012). However, because of the unusually large number of respondents who were allocated to each interviewer on the UKHLS, these ICCs result in high estimated design effects of 3.3 and 2.5 for Q1 and Q2 respectively. Design effects were calculated by using equation (1) with an average cluster size m of 58 for Q1 and 53 for Q2. These represent substantial reductions in precision, indicating that the variance of these estimates is approximately 2–3 times greater than they would be if the interviewer effect were zero. Taking the square root of the design effect gives the inflation factors for the variance of the estimated means, which are 1.8 for Q1 and 1.6 for Q2. Model 1 also shows that there is variability in the magnitude of the residual level 1 variance across interviewers (0.112 and 0.033 for Q1 and Q2).

Table 2 presents the model 2 results for variables Q1 and Q2. Accounting for sample composition differences in model 2 leads to only small changes in the estimated population-averaged ICCs and level 1 residual variances for each question.

Table 1. Model 1 mixed effects location–scale model results for Q1, 'get along with neighbours', and Q2, 'influence politics'†

	Results for Q1				*Results for Q2*			
	Coefficient	*Standard deviation*	*2.5%*	*97.5%*	*Coefficient*	*Standard deviation*	*2.5%*	*97.5%*
Fixed effects								
Location equation, β_0 (intercept)	*1.277*	0.012	1.253	1.300	*3.024*	0.043	2.939	3.110
Scale equation, α_0 (intercept)	*−0.754*	0.023	−0.800	−0.708	*2.135*	0.017	2.102	2.169
Random effects								
$\sigma^2_{u[1]}$ (location: interviewer variance)	*0.024*	0.003	0.018	0.030	*0.266*	0.041	0.192	0.354
$\sigma^2_{u[2]}$ (scale: interviewer variance)	*0.112*	0.014	0.088	0.141	*0.033*	0.006	0.023	0.046
$\sigma_{u[1]u[2]}$ (interviewer cross-equation covariance)	*0.036*	0.005	0.027	0.047	*0.061*	0.011	0.041	0.085
$\sigma^2_{v[1]}$ (location: area variance)	*0.056*	0.004	0.048	0.064	*0.620*	0.064	0.501	0.750
ρ_u (population-average conditional interviewer ICC)	0.041				0.028			

†UKHLS wave 3; Q1 sample size, 303 interviewers, 3473 areas and 17471 respondents; Q2 sample size, 300 interviewers, 3390 areas and 16046 respondents. Q1 DIC = 37829; Q2 DIC = 80773. 'Significant' values are in italics.

Table 2. Model 2 mixed effects location–scale model results for Q1, 'get along with neighbours' and Q2, 'influence politics'†

	Results for Q1				*Results for Q2*			
	Mean	*Standard deviation*	*2.5%*	*97.5%*	*Mean*	*Standard deviation*	*2.5%*	*97.5%*
Fixed effects								
Location equation								
β_0 (intercept)	*1.265*	0.016	1.234	1.297	*2.846*	0.067	2.714	2.976
β_1 (respondent: male)	*0.041*	0.010	0.020	0.061	*−0.103*	0.047	−0.193	−0.011
β_2 (respondent: age)	*−0.042*	0.006	−0.053	−0.031	*0.216*	0.025	0.168	0.265
β_3 (area: ethnic diversity)	0.037	0.056	−0.074	0.146	*1.121*	0.241	0.651	1.597
β_4 (area: socio-economic disadvantage)	*0.126*	0.007	0.112	0.140	*−0.134*	0.032	−0.197	−0.072
β_5 (area: urbanicity)	*0.076*	0.011	0.054	0.098	0.068	0.050	−0.029	0.165
β_6 (area: transitory population)	0.010	0.007	−0.005	0.025	0.037	0.032	−0.027	0.100
β_7 (area: age + housing structure)	−0.030	0.008	−0.045	−0.014	*−0.068*	0.034	−0.135	−0.001
Scale equation								
α_0 (intercept)	*−0.755*	0.023	−0.801	−0.709	*2.127*	0.017	2.094	2.160
Random effects								
$\sigma^2_{u[1]}$ (location: interviewer variance)	*0.019*	0.003	0.014	0.024	*0.258*	0.041	0.185	0.345
$\sigma^2_{u[2]}$ (scale: interviewer variance)	*0.112*	0.014	0.087	0.141	*0.033*	0.006	0.023	0.046
$\sigma_{u[1]u[2]}$ (interviewer cross-equation covariance)	*0.031*	0.005	0.023	0.041	*0.062*	0.011	0.041	0.086
$\sigma^2_{v[1]}$ (location: area variance)	*0.033*	0.003	0.026	0.040	*0.582*	0.064	0.461	0.712
ρ_u (population average conditional interviewer ICC)	0.035				0.028			

†UKHLS wave 3; Q1 sample size, 303 interviewers, 3473 areas and 17471 respondents; Q2 sample size, 300 interviewers, 3390 areas and 16046 respondents; Q1 DIC = 37514; Q2 DIC = 80646. 'Significant' value are in italics.

To provide a more concrete picture of the extent of the variability across interviewers, Fig. 2 plots the sample-corrected interviewer-specific ICCs from model 2 for each interviewer, along with 95% credible intervals and the population-average ICC. Interviewers are ranked from lowest (left) to highest (right) ICC. Across both items it is clear that there is a substantial minority of interviewers with a larger-than-normal correlation between respondents' answers (reaching a maximum of 0.07 for Q1 and 0.04 for Q2). A second group of interviewers has noticeably less similar responses (reaching a minimum of below 0.02 for each question).

Furthermore, the significant positive covariance terms that are reported in Table 2 mean that the level 1 residual variance is higher among interviewers who also have a higher-than-average intercept residual. This covariance may, in part, be an artefact of the scales on which these variables are measured, creating 'floor' effects. That is to say, if responses across all interviewers are low on the response scale, as here, then we would expect interviewers with higher means to have larger variances. As we move from the bottom towards the middle of the response scale, the mean by definition increases, but the variance also rises because more response options are available for respondents to choose from.

Table 3 presents the model 3 results for variables Q1 and Q2. Model 3 adds the interviewer characteristics into the fixed and random parts of the model. Considering the coefficient estimates for the five-point Likert scale item (Q1) first, we find moderate evidence that the mean

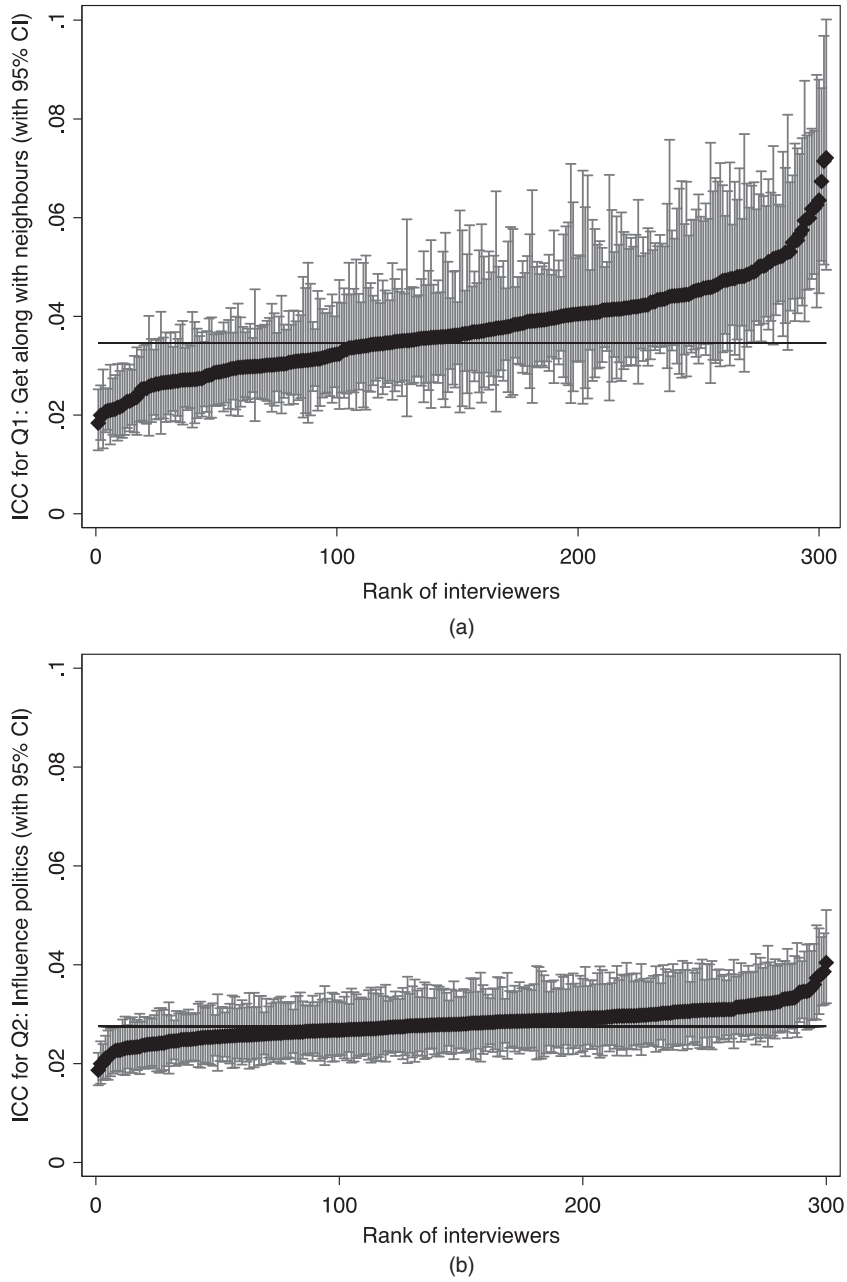

Fig. 2. Interviewer-specific ICCs from model 2 (Table 2) for (a) Q1, 'get along with neighbours', and (b) Q2 'influence politics' (——, population-averaged ICC)

of respondents' answers is influenced by interviewers' views about surveys, with systematically lower mean estimates among interviewers who believe that surveys are generally conducted responsibly, and higher means from interviewers who believe that surveys are mostly correct. No other interviewer variables have a credible interval that excludes zero in the location equation. Turning to the residual variance equation, some interviewer characteristics have significant

Table 3. Model 3 mixed effects location–scale model results for Q1, 'get along with neighbours', and Q2, 'influence politics'†

	Results for Q1				*Results for Q2*			
	Mean	*Standard deviation*	*2.5%*	*97.5%*	*Mean*	*Standard deviation*	*2.5%*	*97.5%*
Fixed effects								
Location equation								
β_0 (intercept)	*1.268*	0.054	1.164	1.373	*2.845*	0.214	2.424	3.263
β_1 (respondent: male)	*0.040*	0.011	0.019	0.061	*−0.097*	0.047	−0.189	−0.006
β_2 (respondent: age)	*−0.042*	0.006	−0.053	−0.031	*0.232*	0.025	0.183	0.281
β_3 (area: ethnic diversity)	0.034	0.057	−0.077	0.145	*1.123*	0.245	0.639	1.602
β_4 (area: socio-economic disadvantage)	*0.126*	0.007	0.112	0.140	*−0.144*	0.032	−0.207	−0.082
β_5 (area: urbanicity)	*0.076*	0.012	0.053	0.098	0.066	0.049	−0.031	0.163
β_6 (area: transitory population)	0.011	0.007	−0.003	0.026	0.034	0.032	−0.027	0.097
β_7 (area: age + housing structure)	*−0.029*	0.008	−0.045	−0.013	−0.066	0.034	−0.133	0.001
β_8 (interviewer: male)	0.016	0.023	−0.029	0.062	*0.212*	0.088	0.042	0.389
β_9 (interviewer: age)	0.018	0.013	−0.007	0.042	−0.061	0.048	−0.157	0.034
β_{10} (interviewer: worked on another survey)	0.001	0.022	−0.042	0.045	0.113	0.086	−0.057	0.283
β_{11} (interviewer: non-survey interviewing)	−0.013	0.023	−0.057	0.032	0.012	0.088	−0.162	0.184
β_{12} (interviewer: public interaction	−0.004	0.027	−0.057	0.047	0.006	0.106	−0.199	0.216
β_{13} (interviewer: survey participation self-interest)	0.026	0.022	−0.017	0.070	−0.008	0.088	−0.177	0.167
β_{14} (interviewer: surveys conducted responsibly)	*−0.127*	0.044	−0.213	−0.042	0.170	0.169	−0.152	0.506
β_{15} (interviewer: surveys correct)	*0.106*	0.041	0.023	0.185	*−0.334*	0.167	−0.667	−0.006
β_{16} (interviewer: agreeableness)	0.005	0.012	−0.019	0.029	0.045	0.047	−0.046	0.136
β_{17} (interviewer: conscientiousness)	0.001	0.012	−0.021	0.025	*0.115*	0.047	0.024	0.206
β_{18} (interviewer: extravert)	0.005	0.012	−0.019	0.028	−0.020	0.046	−0.110	0.070
β_{19} (interviewer: neuroticism)	0.014	0.012	−0.011	0.037	0.008	0.047	−0.083	0.101
β_{20} (interviewer: openness)	0.006	0.012	−0.018	0.029	*0.094*	0.047	0.002	0.185
Scale equation								
α_0 (intercept)	*−0.701*	0.112	−0.915	−0.466	*2.191*	0.084	2.015	2.349
α_1 (respondent: male)	0.004	0.023	−0.041	0.049	*0.050*	0.024	0.004	0.096
α_2 (respondent: age)	*−0.057*	0.012	−0.080	−0.033	*0.094*	0.013	0.069	0.119
α_3 (interviewer: male)	0.090	0.051	−0.009	0.191	0.003	0.035	−0.065	0.070
α_4 (interviewer: age)	0.010	0.027	−0.044	0.062	−0.021	0.019	−0.059	0.015
α_5 (interviewer: worked on another survey)	*0.097*	0.048	0.005	0.192	*0.069*	0.033	0.004	0.134
α_6 (interviewer: worked in public engagement)	−0.016	0.049	−0.111	0.080	−0.040	0.035	−0.111	0.027
α_7 (interviewer: conducted cold calls)	−0.040	0.057	−0.155	0.073	0.021	0.041	−0.059	0.102
α_8 (interviewer: survey participation self-interest)	0.043	0.047	−0.048	0.135	−0.049	0.035	−0.116	0.021
α_9 (interviewer: surveys conducted responsibly)	*−0.269*	0.097	−0.457	−0.077	0.010	0.070	−0.130	0.148
α_{10} (interviewer: surveys correct)	0.125	0.085	−0.040	0.294	−0.122	0.067	−0.257	0.008
α_{11} (interviewer: agreeableness)	0.012	0.025	−0.039	0.062	0.026	0.018	−0.010	0.062
α_{12} (interviewer: conscientiousness)	0.035	0.025	−0.013	0.084	*0.045*	0.019	0.008	0.081
α_{13} (interviewer: extravert)	*0.058*	0.026	0.005	0.106	−0.016	0.018	−0.050	0.020
α_{14} (interviewer: neuroticism)	0.003	0.026	−0.049	0.054	−0.001	0.019	−0.038	0.037
α_{15} (interviewer: openness)	0.014	0.025	−0.035	0.065	0.026	0.018	−0.010	0.062

(*continued*)

Table 3 (*continued*)

	Results for Q1				*Results for Q2*			
	Mean	*Standard deviation*	*2.5%*	*97.5%*	*Mean*	*Standard deviation*	*2.5%*	*97.5%*
Random effects								
$\sigma^2_{u[1]}$ (location: interviewer variance)	*0.019*	0.003	0.014	0.025	*0.237*	0.040	0.167	0.321
$\sigma^2_{u[2]}$ (scale: interviewer variance)	*0.103*	0.013	0.079	0.131	*0.029*	0.005	0.020	0.041
$\sigma_{u[1]u[2]}$ (interviewer cross-equation covariance)	*0.030*	0.005	0.021	0.039	*0.050*	0.011	0.030	0.073
$\sigma^2_{v[1]}$ (location: area variance)	*0.033*	0.003	0.026	0.040	*0.588*	0.064	0.463	0.718
ρ_u (population-average conditional interviewer ICC)	0.035				0.025			

†UKHLS wave 3; Q1 sample size, 303 interviewers, 3473 areas and 17471 respondents; Q2 sample size, 300 interviewers, 3390 areas and 16046 respondents; Q1 DIC = 37498; Q2 DIC = 80590. 'Significant' values are in italics.

effects. This demonstrates the utility of this modelling approach; we detect significant associations between interviewer characteristics and response variance, which would be missed by using the standard random-intercept model.

Interviewers who have prior experience of working on other surveys show a *larger* residual error at the respondent level: an effect which is in line with the results of existing studies (Davis and Scott, 1995; O'Muircheairtagh and Campanelli, 1998; Brunton-Smith *et al.*, 2012). The residual error is also larger among interviewers who are higher on the extraversion dimension of the 'big five' personality inventory, which accords with theoretical expectations; interviewers who are higher on extraversion should be more likely to adopt a more conversational interviewing style. In contrast, the residual error is *lower* among those interviewers who believe that surveys are generally conducted in a responsible way. This association also confirms our *a priori* expectations, with those interviewers who place greater weight on the value of survey research being more likely to stick to standardized interviewing protocols and, therefore, produce less variable responses.

To give some idea of the magnitude of these effects we can take expectations from the model for particular sets of interviewer characteristics. For example, an interviewer, with mean scores on the personality dimensions, who has worked on the UKHLS only and who does not believe that surveys are conducted in a responsible way has an expected ICC of 0.029. If we take an interviewer who shares all these characteristics but believes that surveys are conducted responsibly, the estimated ICC is 0.037. Similarly, an interviewer who has experience of working on another survey has an estimated ICC of 0.027, and an interviewer identified as 1 standard deviation below the average in levels of extraversion has an estimated ICC of 0.031. Although these are small in absolute magnitude, as we saw earlier, differences in the ICC can have a substantial influence on the precision of an estimator when the number of respondents who are interviewed by each interviewer is large.

Turning to the 11-point scale (Q2), the location equation shows that respondents who were interviewed by a male interviewer were more likely to report that they believe that they can influence political decisions, as were respondents whose interviewers scored higher on the conscientiousness and openness personality dimensions. Lower scores were evident among respondents who were interviewed by someone who says that surveys are generally correct. Interviewer gender

Table 4. Model 2 mixed effects location–scale model results for Q3, 'self-completion—belong to neighbourhood'†

	Mean	*Standard deviation*	*2.5%*	*97.5%*
Fixed effects				
Location equation				
β_0 (intercept)	*2.541*	0.019	2.505	2.577
β_1 (respondent: male)	*−0.129*	0.014	−0.156	−0.102
β_2 (respondent: age)	*0.208*	0.007	0.193	0.223
β_3 (area: ethnic diversity)	*0.207*	0.069	0.071	0.345
β_4 (area: socio-economic disadvantage)	*−0.063*	0.009	−0.081	−0.045
β_5 (area: urbanicity)	*−0.114*	0.014	−0.142	−0.086
β_6 (area: transitory population)	−0.006	0.009	−0.024	0.013
β_7 (area: age + housing structure)	*0.032*	0.010	0.013	0.051
Scale equation				
α_0 (intercept)	−0.305	0.015	−0.335	−0.274
Random effects				
$\sigma^2_{u[1]}$ (location: interviewer variance)	*0.013*	0.002	0.009	0.018
$\sigma^2_{u[2]}$ (scale: interviewer variance)	*0.018*	0.004	0.011	0.027
$\sigma_{u[1]u[2]}$ (interviewer cross-equation covariance)	*−0.005*	0.002	−0.010	−0.001
$\sigma^2_{v[1]}$ (location: area variance)	*0.039*	0.005	0.029	0.049
ρ_u (population-average conditional interviewer ICC)	0.016			

†Sample size 302 interviewers, 3383 areas and 15913 respondents; Q3 DIC = 41161. 'Significant' values are in italics.

has emerged as a significant predictor of mean responses on many items in existing studies, although the pattern and magnitude of this effect seems to be item specific (O'Muircheartaigh and Campanelli, 1998). Interviewer characteristics also directly affect the level 1 residual variance. Like for Q1, the residual error is larger among interviewers who have worked on another survey. The residual error is also larger among interviewers who are identified as more conscientious.

Because of the non-random allocation of respondents to interviewers in the UKHLS, it is possible that variability in the magnitude of the ICC across interviewers on these two items may be due to differences in the composition of areas and/or differential non-response across interviewer assignments. To assess this possibility, we fit model 2 to item Q3, which was included in the self-completion questionnaire that was administered as an adjunct to the main interviewer-administered questionnaire. We use the unconditional estimate of the between-interviewer variability from model 2 because this will yield the upper bound of any such potential effect. If the patterns of variance across interviewers that we have observed on items Q1 and Q2 is a reflection of area or non-response confounding, we should expect to see approximately the same between-interviewer variability in the self-completion item. The results are presented in Table 4.

Consistent with the interpretation of our results as resulting from the behaviour of interviewers, Table 4 shows a noticeably smaller interviewer population-averaged ICC (0.016), although we still observe a moderate variance associated with area clustering of 0.039. More importantly, we see almost no variability in the magnitude of the ICC across interviewers (Fig. 3). Because Q3 is self-completion we should not see any influence of interviewers. The significant interviewer variability in the location equation therefore probably reflects differential sample composition

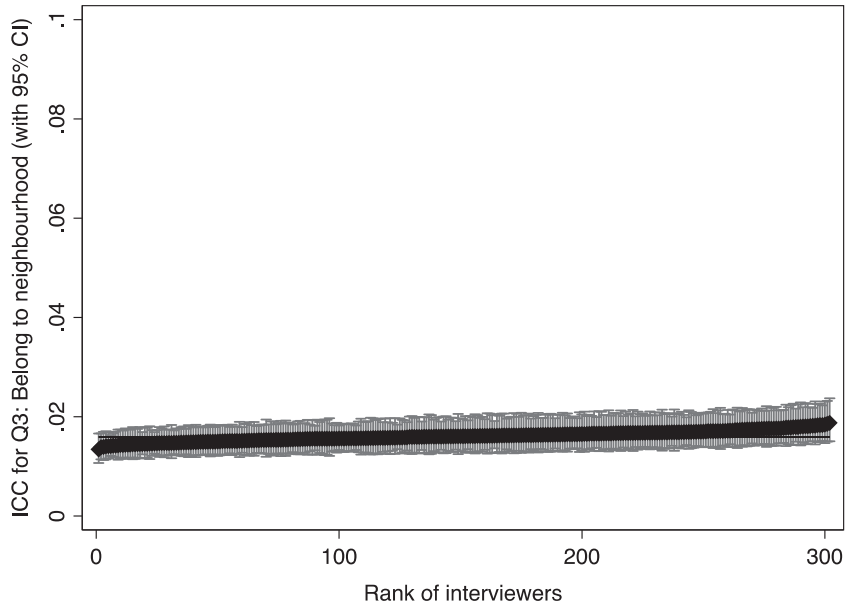

Fig. 3. Interviewer-specific ICCs from model 2 (Table 4) for Q3, 'self-completion—belong to neighbourhood' (——, population-averaged ICC)

across interviewers, although it might also arise from interviewers assisting some respondents to complete the paper questionnaire.

7. Discussion

Survey methodologists have demonstrated that interviewers can substantially reduce the precision of survey parameter estimates through a combination of idiosyncratic behaviours, personal characteristics and dispositions (Hox, 1994; O'Muircheartaigh and Campanelli, 1998; Bailar *et al*., 1977; Finkel *et al*., 1991; Hughes *et al*., 2002). When the number of respondents who are assigned to interviewers is large, standard errors can be inflated by factors of as high as 2, or above. Another way of putting this is that the effective analytical sample size can, in extreme cases, be halved. Even with more standard assignment sizes of around 20 respondents, interviewer ICCs of only 0.03 will inflate standard errors by a factor of approximately 60%. Given the high and increasing unit cost of face-to-face interviews, it is surprising that comparatively little attention has been paid to identifying, and finding ways of reducing, this large and potentially controllable source of survey error.

In this paper we have described a new and more flexible approach than is currently available to detecting and explaining interviewer effects, namely a mixed effects location–scale model. The key benefits of this model are that interviewers can influence variability in respondent level survey responses, on top of any effect that they have on outcome means. The exact mechanism through which interviewer influence comes about remains somewhat opaque but is likely to be due to factors such as failing to follow interview instructions, a tendency to encourage (or discourage) extreme answers, variation in interviewer speed of question delivery, inconsistent use of show cards, and so on. The standard mixed effects random-intercept model does not accommodate the potential for interviewers to influence the variability of the level 1 residual

directly and, as a consequence, may fail to identify important associations with interviewer level characteristics.

We applied the mixed effects location–scale model, with a cross-classified extension, to three attitudinal outcomes from wave 3 of the UKHLS and found notable heterogeneity of variance across interviewers, with some having significantly higher and some significantly lower ICCs than others. At the upper extreme, some interviewers had almost twice the average ICC value for all interviewers. As a result, the design effect for some interviewers will be markedly different from the averages of 3.3 and 2.5 for Q1 and Q2 (estimated from model 1). Across the middle 95% of interviewers this ranges from 2.5 to 4.9 for Q1 and from 2.2 to 2.9 for Q2 (assuming average cluster sizes of 58 and 53 respectively). This approach is therefore of potential value in identifying interviewers who make an unusually large, or indeed small, contribution to the variance of survey parameter estimates. This could form the starting point for targeted training interventions, as well as for developing a better understanding of the behavioural mechanisms which cause interviewer effects in the first place.

We also found systematic differences in interviewer error which were related to observed characteristics of interviewers and, moreover, that these effects differed for the location and scale of the response. That is to say, some interviewer characteristics were associated with variability in the mean of the survey outcome but not with the residual variance, whereas others showed the opposite pattern. Specifically, for the first item that was considered (neighbourhood evaluations) the respondent level residual variance was higher for interviewers with experience of other surveys and lower for interviewers who reported that they believe survey data to be collected responsibly. Interviewers who scored higher on the extraversion dimensions of the 'big five' personality inventory also exhibited significantly more variable responses. Interviewer beliefs about whether survey data are collected responsibly also influenced the mean of respondent answers, as did whether interviewers viewed survey data as generally correct. For the second item (ability to influence politics), four interviewer characteristics—gender, whether they believe that the data are correct and the openness and conscientiousness dimensions of the 'big five'—influenced the mean, whereas differences in the variance were associated with experience of other surveys and conscientiousness. These interviewer characteristic effects can result in substantial differences in the precision of parameter estimates depending on the profile of interviewers. For example, using the parameter estimates from model 3 on item Q1, an interviewer who scored 1 standard deviation below the mean on extraversion, who believes that surveys are conducted responsibly and has worked on the UKHLS only would have an expected design effect of 3.2. In contrast, an interviewer who is 1 standard deviation above the mean on extraversion, who has worked on other surveys and who does not believe that surveys are conducted responsibly has a model-predicted design effect of 2.4. The third item, which was taken from the self-completion schedule of the UKHLS, showed no notable interviewer variance. This served a useful 'sense checking' function as we should not expect to observe interviewer effects on items for which there is little or no interviewer involvement.

Together, these findings suggest some important conclusions relating to interviewer error. First, there is substantial variability across interviewers in the extent to which they affect the precision of survey parameter estimates. Second, interviewer demographic characteristics, survey experience, personality and beliefs about the responses that are provided by participants are significant predictors of this variability. They are, therefore, suggestive of ways in which survey designers might seek to mitigate interviewer-related error through recruitment and training strategies. And, third, interviewer characteristics exert differential effects on the mean and the variance of survey outcomes: a pattern which is dependent on the items considered.

Our primary concern in this paper has been to describe and demonstrate a new methodolog-

ical approach for the study of interviewer effects on the variability of respondents' answers, which is an important though comparatively neglected source of survey error. Although our analyses have produced substantively interesting and meaningful results, our focus on analytical explication has meant that the methodology has been foregrounded at the expense of substantive generality. Further research is required to evaluate how well our findings generalize across a wider range of question types and survey contexts, as well whether and how training interventions might be effective in reducing the kinds of interviewer error that the model identifies.

Acknowledgement

The authors gratefully acknowledge the support of the Economic and Social Research Council through the grant for the National Centre for Research Methods (reference ES/L008351/1).

References

Anderson, B. A., Silver, B. D. and Abramson, P. R. (1988) The effects of race of the interviewer on measures of electoral participation by blacks in SRC national election studies. *Publ. Opin. Q.*, **52**, 53–83.

Bailar, B. A., Bailey, L. and Stevens, J. (1977) Measures of interviewer bias and variance. *J. Marktng Res.*, **24**, 337–343.

Biemer, P. P. and Stokes, S. L. (1985) Optimal design of interviewer variance experiments in complex surveys. *J. Am. Statist. Ass.*, **80**, 158–166.

Brunton-Smith, I. and Sturgis, P. (2011) Do neighbourhoods generate fear of crime?: an empirical test using the British Crime Survey. *Criminology*, **49**, 331–369.

Brunton-Smith, I., Sturgis, P. and Williams, J. (2012) Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance? *Publ. Opin. Q.*, **76**, 265–286.

Burton, J., Knies, G. and Al Baghal, T. (2014) Understanding society: interviewer survey 2014. *User Guide v1.1.* Colchester: UK Household Longitudinal Survey.

Cannell, C. F., Miller, P. V. and Oksenberg, L. (1981) Research on interviewing techniques. In *Sociological Methodology* (ed. S. Leinhardt), pp. 389–437. San Francisco: Jossey-Bass.

Charlton, C. M. J., Michaelides, D. T., Parker, R. M. A., Cameron, B., Szmaragd, C., Yang, H., Zhang, Z., Frazer, A. J., Goldstein, H., Jones, K., Leckie, G., Moreau, L. and Browne, W. J. (2013) StatJR version 1.0. Centre for Multilevel Modelling, University of Bristol and Electronics and Computer Science, University of Southampton. (Available from `http://www.bristol.ac.uk/cmm/software/statjr/`.)

Collins, M. (1980) Interviewer variability: a review of the problem. *J. Markt Res. Soc.*, **22**, 77–95.

Davis, P. and Scott, A. (1995) The effect of interviewer variance on domain comparisons. *Surv. Methodol.*, **21**, 99–106.

Durrant, G. B., Groves, R. M., Staetsky, L. and Steele, F. (2010) Effects of interviewer attitudes and behaviors on refusal in household surveys. *Publ. Opin. Q.*, **74**, 1–36.

Fellegi, I. P. (1964) Response variance and its estimation. *J. Am. Statist. Ass.*, **59**, 1016–1041.

Fellegi, I. P. (1974) An improved method of estimating the correlated response variance. *J. Am. Statist. Ass.*, **69**, 496–501.

Finkel, S. E., Guterbock, T. M. and Borg, M. J. (1991) Race-of-interviewer effects in a preelection poll: Virginia, 1989. *Publ. Opin. Q.*, **55**, 313–330.

Gelman, A. (2007) Struggles with survey weighting and regression modelling. *Statist. Sci.*, **22**, 153–164.

Goldberg, L. (1990) An alternative "description of personality": the big-Five factor structure. *J. Personlty Socl Psychol.*, **59**, 1216–1229.

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

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961) Measurement errors in censuses and surveys. *Bull. Int. Statist. Inst.*, **38**, 359–374.

Hedeker, D., Mermelstein, R. J. and Demirtas, H. (2008) An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, **64**, 627–634.

Hedeker, D. and Nordgren, R. (2013) MIXREGLS: a program for mixed-effects location scale analysis. *J. Statist. Softwr.*, **52**, 1–38.

Hirschman, A. O. (1964) The paternity of an index. *Am. Econ. Rev.*, **54**, 761.

Hox, J. J. (1994) Hierarchical regression models for interviewer and respondent effects. *Sociol. Meth. Res.*, **22**, 300–318.

Hox, J. J., de Leuww, E. D. and Kreft, I. I. G. (1991) The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys* (eds P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman), pp. 439–462. New York: Wiley.
Huddy, L., Billig, J., Bracciodieta, J., Hoeffer, L., Moynihan, P. and Puglian, P. (1997) The effect of interviewer gender on the survey response. *Polit. Behav.*, **19**, 197–220.
Hughes, A., Chromy, J., Giacoletti, K. and Odom, D. (2002) Impact of interviewer experience on respondent reports of substance use. In *Redesigning an Ongoing National Household Survey: Methodological Issues, Substance Abuse and Mental Health Services Administration* (eds J. Gfroerer, J. Eyerman and J. Chromy), pp. 161–184. Rockville: Office of Applied Studies.
Jäckle, A., Lynn, P., Sinibaldi, J. and Tipping, S. (2013) The effect of interviewer experience, attitudes, personality and skills on respondent cooperation with face-to-face surveys. *Surv. Res. Meth.*, **7**, 1–15.
Kane, E. W. and Macaulay, L. J. (1993) Interviewer gender and gender attitudes. *Publ. Opin. Q.*, **57**, 1–28.
Kish, L. (1962) Studies of interviewer variance for attitudinal variables. *J. Am. Statist. Ass.*, **57**, 92–115.
Knies, G. (2014) Understanding society: the UK Household Longitudinal Study Waves 1-4. *User Manual*. Institute for Social and Economic Research, Colchester.
Leckie, G., French, R., Charlton, C. and Browne, W. (2014) Modeling heterogeneous variance-covariance components in two-level models. *J. Educ. Behav. Statist.*, **39**, 307–332.
Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012) *The BUGS Book: a Practical Introduction to Bayesian Analysis*. Boca Raton: Chapman and Hall–CRC.
Mangione, T., Fowler, F. J. and Louis, T. A. (1992) Question characteristics and interviewer effects. *J. Off. Statist.*, **8**, 293–307.
Martin, D. (2001) *Geography for the 2001 Census in England and Wales*. London: Office for National Statistics.
O'Muircheartaigh, C. (1976) Response errors in an attitudinal sample survey. *Qual. Quant.*, **26**, 97–115.
O'Muircheartaigh, C. and Campanelli, P. (1998) The relative impact of interviewer effects and sample design effects on survey precision. *J. R. Statist. Soc.* A, **161**, 63–77.
Pickery, J., Loosveldt, G. and Carton, A. (2001) The effects of interviewer and respondent characteristics on response behaviour in panel surveys: a multilevel approach. *Sociol. Meth. Res.*, **29**, 509–523.
Rao, J. N. K., Verret, F. and Hidiroglou, M. A. (2013) A weighted composite likelihood approach to inference for two-level models from survey data. *Surv. Methodol.*, **39**, 263–282.
Schaeffer, N. C. (1980) Evaluating race of interviewer effects in a national survey. *Sociol. Meth. Res.*, **8**, 400–419.
Schaeffer, N. C., Dykema, J. and Maynard, D. W. (2010) Interviewers and interviewing. In *Handbook of Survey Research*, 2nd edn (eds P. V. Marsden and J. D. Wright). Bingley: Emerald.
Schnell, R. and Kreuter, F. (2005) Separating interviewer and sampling point effects. *J. Off. Statist.*, **21**, 389–410.
Scott, A. and Jessop, C. (2013) UK Household Longitudinal Study (UKHLS) Wave 3. *Technical Report*. NatCen, London.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Sturgis, P. and Luff, R. (2015) Audio-recording of open-ended survey questions: a solution to the problem of interviewer transcription? In *Survey Measurements: Techniques, Data Quality and Sources of Error* (ed. U. Engel), pp. 42–57. Chicago: University of Chicago Press.
van Tilburg, T. (1998) Interviewer effects in the measurement of personal network size: a nonexperimental study. *Sociol. Meth. Res.*, **26**, 300–328.
Turner, M., Sturgis, P., Martin, D. and Skinner, C. (2014) Can interviewer personality, attitudes and experience explain the design effect in face-to-face surveys? In *Improving Survey Methods: Lessons from Recent Research* (eds U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis). Abingdon: Routledge.
West, B. T. and Elliott, M. R. (2014) Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Surv. Methodol.*, **40**, 163–188.
West, B. T., Kreuter, F. and Jaenichen, U. (2013) Interviewer effects in face-to-face surveys: a function of sampling, measurement error or nonresponse? *J. Off. Statist.*, **29**, 277–297.
West, B. T. and Olson, K. (2010) How much of interviewer variance is really nonresponse error variance? *Publ. Opin. Q.*, **74**, 1004–1026.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online appendix: Running the Stat-JR cross-classified mixed-effects location scale model template'.