Bayesian Analysis of Between-Group Differences in Variance Components in Hierarchical Generalized Linear Models

Brady T. West[1]
[1] Michigan Program in Survey Methodology, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48104

Abstract
Frequentist approaches to making inferences about the variances of random cluster effects in hierarchical generalized linear models (HGLMs) for non-normal variables have several limitations. These include reliance on asymptotic theory, questionable properties of classical likelihood ratio tests when pseudo-likelihood methods are used for estimation, and a failure to account for uncertainty in the estimation of features of prior distributions for model parameters. This paper compares and contrasts alternative approaches to making a specific type of inference about the variance components in an HGLM, focusing on the difference in variance components between two independent groups of clusters. A Bayesian approach to making inferences about these types of differences is proposed that circumvents many of the problems associated with alternative frequentist approaches. The Bayesian approach and alternative frequentist approaches are applied to an analysis of real survey data collected in the Continuous National Survey of Family Growth (NSFG). The primary analytic question of interest concerns differences in the variances of random interviewer effects between two independent groups of interviewers, which may indicate that particular subsets of interviewers are having adverse effects on the quality of the survey data. Inferences regarding differences in interviewer variance components are shown to vary depending on the approach taken, with significant differences suggested by problematic frequentist analyses no longer evident when applying more appropriate Bayesian analysis methods.

Key Words: Bayesian Analysis, Variance Components, Interviewer Variance, Hierarchical Generalized Linear Models

1. Introduction

Hierarchical Generalized Linear Models (HGLMs), which are also referred to as Generalized Linear Mixed Models (GLMMs), are flexible analytic tools used in a variety of scientific fields to analyze clustered[1] data sets, where observations on a non-normal response variable of interest within a cluster cannot be considered independent. This paper considers alternative approaches to making inferences about the parameters in a specific class of HGLMs, where the variances of random cluster effects for two independent groups of clusters defined by a known cluster-level covariate may not be equal. The paper considers an application of the approaches to a specific problem in the field of survey methodology, where the objective of an analysis is to compare two groups of interviewers in terms of their variance components for a non-normal survey variable of interest. If one group of interviewers has a significantly higher variance component than another, that group may be having an adverse effect on the quality of the survey estimates, and intervention is needed. More generally, the approaches discussed in this paper would allow one to make inferences about differences in these variance components in a variety of settings.

Frequentist approaches to estimation of HGLMs rely on various numerical or theoretical approaches to approximating complicated likelihood functions, especially for models involving

[1] 'Clustered' data sets include longitudinal data sets for the purposes of this paper.

complex random effects structures (e.g., Faraway, 2006, Chapter 10; Molenberghs and Verbeke, 2005). In general, inferences based on these approximate likelihood-based approaches, such as residual pseudo-likelihood, penalized quasi-likelihood, and maximum likelihood based on a Laplace approximation, have the same drawback as HLMs for normal outcomes in that they fail to account for the uncertainty in estimating features of prior distributions for the model parameters (Carlin and Louis, 2009, p. 335-336). In addition, frequentist approaches to testing hypotheses about fixed effects or covariance parameters in HGLMs and making inferences about the parameters rely on asymptotic theory and asymptotic results (Zhang and Li, 2010). Molenberghs and Verbeke (2005, p. 277) argue that likelihood ratio tests should not even be used to test hypotheses when models are fitted using pseudo-likelihood methods. Furthermore, the number of clusters under study may be fairly small in practice, making inferences or tests of hypotheses concerning between-cluster covariance parameters based on asymptotic theory invalid. Approximate maximum likelihood estimation methods can also lead to invalid (i.e., negative) estimates of variance components in these models. Bayesian methods for making inferences about the parameters in HGLMs can provide an attractive solution to these various problems, and this paper considers such methods.

In general, HGLMs allow analysts to fit models with a variety of covariance structures for the random effects and/or random errors in the models, although more complex structures can lead to computational difficulties. For example, the simplest HGLM that a frequentist can fit to a non-normal dependent variable *Y* has the following general form:

$$g(E[y_{ij} \mid u_i]) = \beta_0 + u_i,\ u_i \sim N(0, \tau^2) \tag{1}$$

In (1), *i* is an index for a randomly selected cluster of correlated observations on the non-normal dependent variable *Y*, and *j* is an index for individuals nested within each cluster. The dependent variable *Y* with measured values y_{ij} has an expectation conditional on the random cluster effects u_i that is defined by an assumed distribution [e.g., Bernoulli, with conditional expectation $E[y_{ij} \mid u_i] = \pi_{ij}$, with $\pi_{ij} = P(y_{ij} = 1 \mid u_i)$], and *g*(.) is a link function specific to a generalized linear model (e.g., the logit link for a logistic regression model, or the log link for Poisson regression model). The fixed effect β_0 is the intercept in the model, with $g^{-1}(\beta_0)$ representing the overall mean for *Y*. The random effects are often assumed to follow a normal distribution with mean 0 and variance τ^2, although other distributions can be considered.

The variance of the dependent variable *Y* in an HGLM is defined generally as

$$Var[y_{ij} \mid u_i] = \phi w_{ij}^{-1} v(E[y_{ij} \mid u_i]), \tag{2}$$

where ϕ is a parameter allowing for over-dispersion relative to the assumed distribution, w_{ij} is a pre-specified weight specific to the model being fitted (in many cases equal to 1), and *v*(.) is the variance function defined by the assumed distribution (e.g., $v(E[y_{ij} \mid u_i]) = E[y_{ij} \mid u_i]$ for the Poisson model).

The standard frequentist approach to making inference about τ^2 in (1) (and thus the need for the random effects in the HGLM) would involve testing the (questionable) null hypothesis that $\tau^2 = 0$ using a likelihood ratio test statistic that asymptotically follows a mixture of two chi-square distributions with degrees of freedom equal to 0 and 1 and equal weight 0.5 (Zhang and Lin, 2010). In HGLMs, this distribution does not always hold, and other methods may be needed for

making inference regarding the variance components (e.g., Sinha, 2009). Zhang and Lin (2010) review developments in likelihood ratio tests and score tests for variance components in GLMMs more generally.

This paper considers frequentist and Bayesian methods for making inferences about a specific type of covariance structure for the random effects in an HGLM, or one where two different groups of clusters have difference variance components. An example of an HGLM of this type follows:

$$\begin{aligned} &g(E[y_{ij} \mid u_i]) = \beta_0 + \beta_1 I(Group = 1)_i + u_{i(1)} I(Group = 1)_i + u_{i(2)} I(Group = 2)_i \\ &u_{i(1)} \sim N(0, \tau_1^2),\ u_{i(2)} \sim N(0, \tau_2^2) \end{aligned} \tag{3}$$

In this case, there are known binary indicators [e.g., $I(Group = 1)_i$] for two groups of clusters available in the data. The random cluster effects from Group 1 are assumed to follow a normal distribution with mean 0 and variance τ_1^2, while the random cluster effects from Group 2 are assumed to follow a normal distribution with mean 0 and variance τ_2^2. As before, other distributions may be posited for the random effects; the key idea is that random effects for different groups of clusters have different variances. The fixed effect parameter β_1 in (3) represents a fixed effect of Group 1 on the outcome relative to Group 2 in the HGLM.

Analytic interest lies in the magnitude of the difference in the two variance components. A standard frequentist approach to making inference about the difference in the variance components would involve testing the null hypothesis that $\tau_1^2 = \tau_2^2$, versus the alternative hypothesis that $\tau_1^2 \neq \tau_2^2$. Conceptually, this is a simple hypothesis test to perform using frequentist methods, as the null hypothesis defines an equality constraint rather than setting a parameter to a value on the boundary of a parameter space, as in (1). The model under the null hypothesis is nested within the model under the alternative hypothesis, where $\tau_2^2 = \tau_1^2 + k$. The null hypothesis can thus be rewritten as $k = 0$, versus the alternative that $k \neq 0$. A standard likelihood ratio test for one parameter can therefore be applied to test the null hypothesis in the frequentist setting. However, concern about this test arises when the number of clusters in each group is small and asymptotic results for the likelihood ratio test may no longer hold, or when pseudo-likelihood methods are used to estimate the models being compared (Molenberghs and Verbeke, 2005, p. 277).

This paper considers applications of models having the form in (3) to the context of interviewer variance in survey methodology (e.g., O'Muircheartaigh and Campanelli, 1998; Kish, 1962), where survey respondents nested within interviewers (clusters) may appear to be more similar than respondents with different interviewers. If random subsamples of sample units are assigned to interviewers following an interpenetrated design (Mahalanobis, 1946), this is an unfortunate byproduct of the survey data collection process that can increase the variance in survey estimates of means (e.g., Groves, 2004, p. 364). Significant variance components due to interviewers in (1) may arise due to correlated response deviations introduced by an interviewer (e.g., Biemer and Trewin, 1997) or nonresponse error variance among interviewers (West and Olson, 2010). Survey research organizations train interviewers to eliminate this component of variance, which ideally would result in $\tau^2 = 0$ in (1). In reality, this component of variance can never be equal to 0, but survey managers aim to minimize the component via specialized interviewer training. Many

survey variables are also non-normal in nature (binary indicators, counts, or categorical responses), introducing the need for HGLMs to properly analyze interviewer effects on the data.

Models of the form in (3) can be applied when methodological studies are designed to compare two different groups of interviewers in terms of their variance components. For example, a debate exists in the survey methodology literature regarding whether interviewers should use standardized or conversational interviewing. Proponents of standardized interviewing argue that all interviewers should administer surveys in the exact same way, allowing respondents to interpret questions as they see fit (e.g., Fowler and Mangione, 1990). Other research has shown that more flexible interviewing using a conversational style may increase respondent understanding of survey questions and reduce measurement error (e.g., Schober and Conrad, 1997). To test a hypothesis that one interviewing style results in lower between-interviewer variance, a researcher might randomize interviewers to two groups trained in the two different styles, collect survey data on a variety of variables, and then fit model (3), including indicator variables for the two groups of interviewers. Frequentist approaches relying on asymptotic results for tests comparing the variance components in this setting would be limited, given that the number of interviewers in each group will likely be small.

A Bayesian approach to making inferences about differences in variance components in this context has several attractive features relative to the frequentist approach. The Bayesian approach would not require asymptotic theory or assumed asymptotic distributions for the test statistics computed in the frequentist approach, would account for the uncertainty in estimating features of prior distributions for model parameters, and would allow analysts to construct credible intervals for the difference between the two variance components based on draws from a posterior distribution for the two variance components (treating the fixed effects and any additional error variances allowing for possible overdispersion in the non-normal responses as nuisance parameters). This paper compares and contrasts these alternative approaches using real data collected in the Continuous National Survey of Family Growth (NSFG).

2. Methods

2.1 Data

This paper presents analyses of real data collected in Quarter 1 of the NSFG (June 2006 – September 2006). The original design of the NSFG (Groves et al., 2009) called for 16 quarters of data collection from a continuous sample that will be nationally representative when completed in June 2010. The data analyzed in this paper were collected from a national sample of 762 females between the ages of 15 and 44 by 38 female interviewers (with varying sample sizes for each interviewer). Each interviewer has information available on their marital status (21/38 are married), years of interviewing experience (21/38 have years of experience greater than or equal to the median for all interviewers, or 4 years), and other employment status (20/38 are also working another job). Considering one of these interviewer-level variables at a time, the variables will be used to divide the interviewers into two groups (in the absence of an ideal randomized experiment, like that described in the Introduction). Each female respondent has their parity (or count of live births) measured, which will be analyzed as the key survey variable Y (assumed to follow an overdispersed Poisson distribution). The primary analytic question is whether these different groups of female interviewers have significantly different variance components for this particular survey variable.

2.2 Model

The following overdispersed hierarchical Poisson regression model will be fitted to the measured parity data, considering one binary interviewer-level variable at a time:

$$
\begin{aligned}
&y_{ij} \sim Poisson(\theta_i) \\
&\log\theta_i = \beta_0 + \beta_1 I(Group=1)_i + u_{i(1)} I(Group=1)_i + u_{i(2)} I(Group=2)_i + \varepsilon_{ij} \\
&u_{i(1)} \sim N(0,\tau_1^2),\ u_{i(2)} \sim N(0,\tau_2^2),\ \varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)
\end{aligned} \quad (4)
$$

In this model, y_{ij} is the parity for female respondent j collected by interviewer i, $\exp(\beta_0)$ represents the expected parity for Group 2, $\exp(\beta_1)$ represents the expected multiplicative change in parity for Group 1 relative to Group 2, $u_{i(1)}$ is a random effect associated with interviewer i in Group 1, $u_{i(2)}$ is a random effect associated with interviewer i in Group 2, and the random ε_{ij} term allows for overdispersion relative to the standard Poisson distribution (with mean equal to the variance). The parameter σ_ε^2 measures the amount of overdispersion, with $\sigma_\varepsilon^2 = 0$ corresponding to classical Poisson regression.

Alternatively, one can introduce an overdispersion parameter ϕ in this model, such that

$$Var[y_{ij} \mid u_i] = \phi E[y_{ij} \mid u_i]. \quad (5)$$

Estimation of this parameter is common in frequentist approaches to fitting Poisson regression models (e.g., Faraway, 2006), with $\phi = 1$ corresponding to classical Poisson regression. This approach will be contrasted with a Bayesian approach to fitting the model defined in (4).

The fully Bayesian approach to fitting this model will consider noninformative prior distributions for the two fixed effects and the three variance parameters in (4). As mentioned earlier, the frequentist approach does not incorporate the uncertainty in the estimation of these parameters, which is an advantage of the Bayesian approach. Three models of the form defined in (4) above will be fitted using frequentist and Bayesian approaches, considering each of the three interviewer-level indicators separately.

2.3 Frequentist Approach

The parameters in the model defined in (4) will be estimated using residual pseudo-likelihood (RPL) estimation, as implemented in the GLIMMIX procedure in the SAS/STAT software (SAS, 2010). This estimation method was selected instead of maximum likelihood using Laplace approximation or adaptive quadrature for two main reasons. First, previous work has found favorable simulation results for this approach indicating nearly unbiased estimation of the variance components in a GLMM (Pinheiro and Chao, 2006), similar to the case of REML estimation in an HLM for normal data. In addition, this is a more flexible estimation method that allows for the incorporation of overdispersion parameters (e.g., ϕ in (5) above) into the model (referred to as "R-side covariance structures" in the online GLIMMIX documentation). Estimated standard errors will be computed for the parameter estimates using Taylor series linearization methods, which do not account for the uncertainty in estimating the variances of the random effects and any residual terms (SAS, 2010).

For a given interviewer-level indicator, a standard likelihood ratio test will be used to test the null hypothesis that $\tau_1^2 = \tau_2^2$, versus the alternative hypothesis that $\tau_1^2 \neq \tau_2^2$. This test statistic is computed by fitting a version of (4) with the random effect variance components in the two

groups constrained to be equal (Model 2), and then fitting (4), which will be referred to as Model 1. The positive difference in -2 pseudo-log-likelihood values of these two models is then computed, and referred to a chi-square distribution with one degree of freedom. This is an asymptotic result, however, that may not hold well for this particular context with small counts of clusters (interviewers) in the two groups. Further, use of this test is questionable when pseudo-likelihood methods are used to estimate the two models (Molenberghs and Verbeke, 2005, p. 277). Despite these issues, this test is readily implemented in the GLIMMIX procedure through the COVTEST statement with the HOMOGENEITY option, and we consider this naïve test in the frequentist analyses for comparison with the Bayesian approach only. The relevant code for fitting these models using the GLIMMIX procedure is included in the Appendix.

Sensitivity of the estimates to omission of a fixed group effect will also be considered in the frequentist analyses.

2.4 Bayesian Approach

The Bayesian approach to fitting the HGLM described in (4) will use a Gibbs sampler based on the adaptive rejection sampling methodology (Gilks and Wild, 1992), as implemented in the BUGS (Bayesian Inference using Gibbs Sampling) software[2], to simulate draws from the posterior distribution for the parameters in the general model defined in (4). Diffuse noninformative priors for the fixed effects and the variance parameters will be specified for the simulations, to let the data provide the most information about the posterior distributions of the parameters (although in a continuous survey design like that of the NSFG, information from previous quarters could certainly be used to define more informative prior distributions for the parameters). This approach will enable inferences based on simulated draws from the marginal posterior distributions of the two fixed effect parameters, the three variance parameters, the 38 random interviewer effects, and any functions of these parameters. This study focuses on the marginal posterior distribution of the difference in the random effect variances for two groups of interviewers defined by a known binary interviewer-level factor (e.g., marital status), computed using the simulated draws of the two variance components.

Specifically, the following noninformative prior distributions for these parameters will be used. These prior distributions are selected based on a combination of estimates from initial naïve model fitting and recommendations from Gelman and Hill (2007) and Gelman (2006, Section 7) for proper but noninformative prior distributions for variance parameters in hierarchical models with a reasonably large number (i.e., more than 5) of groups (or interviewers, in the present context):

$$\beta_0 \sim N(0,100),\ \beta_1 \sim N(0,100)$$
$$\tau_1^2 \sim Uniform(0,10),\ \tau_2^2 \sim Uniform(0,10),\ \sigma_\varepsilon^2 \sim Uniform(0,10)$$

The noninformative priors for the fixed effects indicate an expectation that the coefficients will be somewhere in the range (-10, 10), while the non-informative priors for the variance components are uniform distributions on the range (0, 10). Given initial naive estimates of the fixed effects ranging between -1 and 1 and initial estimates of the variance components ranging between 0 and 5, these priors are all fairly diffuse, expressing little prior knowledge about these parameters and letting the available data provide the most information. Prior studies comparing interviewer variance components for similar count variables (or prior quarters in continuous surveys) could also be used in general applications of this technique to specify more informative prior

[2] http://www.mrc-bsu.cam.ac.uk/bugs/welcome.html

distributions. It is also important to note that the BUGS software uses inverse-variances for the normal distribution, meaning that 0.01 and inverses of the variance components will be specified in the normal distribution functions (see the BUGS code in the Appendix).

Inference for the difference in variance components between two groups of interviewers will be based on several thousand draws of the two variance components from the joint posterior distribution estimated using the Gibbs sampler. For each draw d of the two variance components, the difference in the variance components, defined as $\tau_1^{2(d)} - \tau_2^{2(d)}$, will be computed. Inferences will then be based on the marginal distribution of these differences, ignoring the draws of the random interviewer effects and the other nuisance parameters. The median and the 0.025 and 0.975 quantiles (for a 95% credible set) of the simulated differences of the two variance components will be computed based on the effective number of simulation draws of the two variance components from the estimated joint posterior distribution. In each analysis, 5,000 draws from the posterior distribution will be generated using the Gibbs sampler, with 2,500 draws discarded as burn-in draws, and the effective number of simulation draws will be computed by BUGS based on the 2,500 draws (Gelman and Hill, 2007, Chapter 16). If the 95% credible set includes 0, there will be evidence in favor of the two groups having equal variance components. If the 95% credible set does not include 0, there will be evidence in favor of the two groups having different variances, with a positive median suggesting that group 1 has the higher variance component. Inference for the two fixed effects and the overdispersion parameter will follow a similar approach.

Focusing on draws of the two variance components from the full joint posterior distribution (and their differences) and ignoring draws of the random interviewer effects and the other nuisance parameters (i.e., the fixed effects and the overdispersion parameters) has the effect of integrating these other parameters out of the joint posterior distribution. This Bayesian approach therefore provides a convenient methodology for simulating draws from the marginal distribution of a complicated parameter (the difference between the two variance components), which would not be possible in the frequentist approach. Sensitivity of inferences to misspecification of the general model in (4) (e.g., omitting the fixed group effect) will also be examined, as in the frequentist case.

Three (3) Markov chains will be run in parallel in the iterative BUGS Gibbs sampling algorithm to simulate random walks through the space of the joint posterior distribution. Three random draws from the standard normal distribution will be used to define starting values for the three chains for each of the two fixed effects and the 38 random interviewer effects. Three random draws from the UNIFORM(0,1) distribution will be used to define starting values for the three chains for each of the three variance components. The Gelman-Rubin $\hat{R}$ statistic, representing (approximately) the square root of the variance of the mixture of the three chains divided by the average within-chain variance (Gelman and Rubin, 1992), will be used to assess convergence (or mixing) of the chains for each parameter. Values less than 1.1 on this statistic will be considered as evident of convergence of the chains for a given parameter. Posterior draws of the parameters will be pooled from the three chains to generate the final effective sample size of draws used for inferences.

Following Gelman and Hill (2007, Chapter 24), posterior predictive checks of the fit of each model will be conducted as well. Specifically, each model will be checked using discrepancy variables defined by the mean and standard deviation of both observed parity values and predicted parity values based on the simulated draws of the parameters from the joint posterior distribution. For each set of simulated draws of the five parameters and 38 random effects in (4)

in the effective sample of simulations, the full data set of 762 parity values will be simulated based on the model in (4), and the mean and standard deviation of the replicated parity values will be computed. The distributions of these simulated posterior means and standard deviations across the sets of simulated draws will be compared to the observed mean and standard deviation on parity, and posterior predictive p-values for assessing model fit will be estimated as the proportion of replicated means and standard deviations greater than the observed mean and standard deviation (Gelman and Hill, 2007, p. 514, 520-521). Finally, the Decision Information Criterion (DIC) will be computed as an estimate of predictive error for each model, enabling comparisons of the alternative models based on the interviewer groups (with smaller DIC values indicating better-fitting models). R and BUGS code for the analysis is available upon request.

3. Results

3.1 Descriptive Statistics

Table 1 presents descriptive statistics for the interviewers in each of the groups defined by the three binary interviewer-level factors. These descriptive statistics include the number of interviewers in each group (out of 38 total), the mean, standard deviation (SD) and range for the number of cases (sample sizes) assigned to each interviewer, and the range of observed means on the parity variable.

Table 1: Descriptive statistics for interviewers in each group defined by the three binary interviewer-level factors

	Number of Interviewers	Total Sample Size	Mean Sample Size	SD of Sample Sizes	Range of Sample Sizes	Range of Parity Means
Experience						
4+ Years	21	408	19.43	8.81	(6, 38)	(0.70, 2.50)
< 4 Years	17	354	20.82	10.13	(3, 46)	(0.42, 1.96)
Married						
Yes	21	458	21.81	9.31	(3, 46)	(0.42, 2.50)
No	17	304	17.88	9.12	(6, 33)	(0.78, 1.86)
Other Job						
Yes	20	409	20.45	9.09	(6, 46)	(0.42, 2.50)
No	18	353	19.61	9.80	(3, 38)	(0.58, 2.50)

The descriptive statistics in Table 1 indicate substantial variance in the sizes of the samples assigned to the interviewers, with roughly 20 cases assigned to each interviewer on average in each of the groups. A modeling approach treating interviewer effects as fixed would probably not make sense for these data, given the small sample sizes for some of the interviewers (which could lead to unstable estimates for particular interviewers). Instead, a modeling approach that borrows information across interviewers (treating interviewer effects as random) would lead to more stable estimates of means for each interviewer. There is also evidence of a larger range in parity means among married interviewers compared to interviewers who are not married. The two groups defined by experience and other job status are similar in terms of ranges of parity means.

Simple examinations of the distributions of the means of observed parity measures for the interviewers in each group are presented in Figure 1 below, to obtain an initial sense of the magnitude of interviewer variance in each of the groups. Figure 1 presents side-by-side box plots of the interviewer means on the parity variable for each group, with the means weighted by assigned sample sizes, along with the overall distribution of the 762 parity measures in the complete data set.

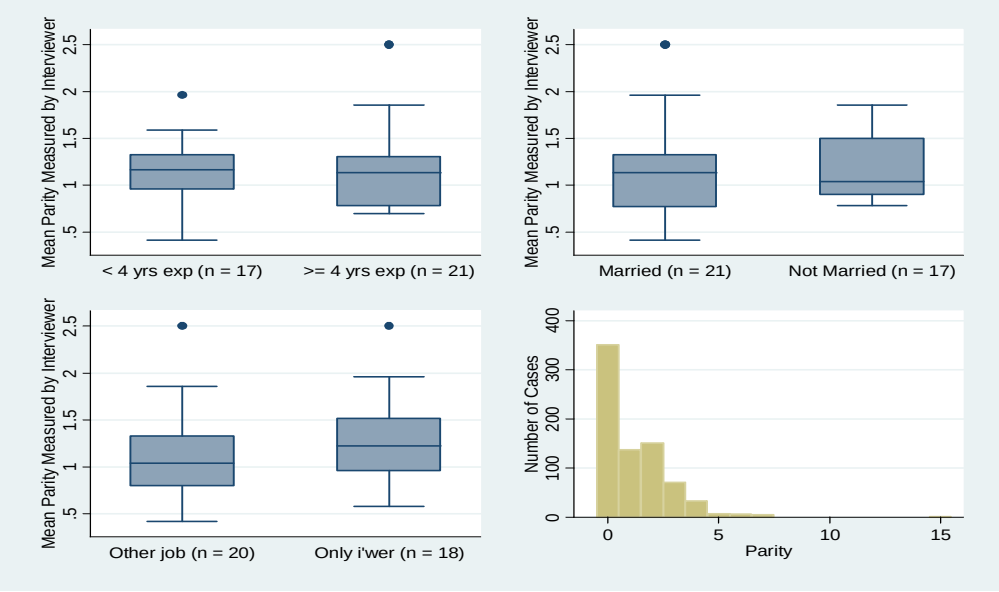

Figure 1: Distributions of observed means on parity for interviewers in each group defined by interviewing experience (less than 4 years vs. 4+ years), marital status (married vs. not married), and other job status (other job vs. only interviewer) with interviewer means weighted by assigned sample size, along with the overall distribution of the parity measures

The distributions of the means of measured parity values for the interviewers in Figure 1 provide an initial sense of groups that tend to differ in terms of the interviewer variance components. The group with more experience appears to have slightly more variance in the means (relative to the group with lower experience), as does the group of interviewers that is married (relative to the interviewers that are not married). The two groups defined by other job status appear to have equal variance in means. The box plots also suggest that the groups do not vary substantially in terms of parity means, which is reassuring (i.e., different groups of interviewers are not significantly impacting the estimate of interest). Finally, the distribution of observed parity values for all 762 respondents has the expected appearance for a variable measuring a count of relatively rare events (live births). The mean of all observed parity values is 1.19, while the variance of all observed values is 2.23. These results suggest that a hierarchical Poisson regression model allowing for overdispersion would probably be reasonable for these data.

3.2 Model Fitting

Table 2 compares estimates of the parameters in the model comparing the two groups of interviewers defined by experience, using the frequentist and Bayesian approaches. Table 2 also presents a naive likelihood ratio test of the null hypothesis that the two experience groups have equal random effect variances, and a 95% posterior credible set for the difference in the two variance components based on the Bayesian approach.

Table 2: Comparisons of frequentist and Bayesian inferences for high vs. low interviewer experience

	Frequentist Analysis*				Bayesian Analysis**			
Parameter	RPL Estimate	LSE	95% CI LL	95% CI UL	Posterior Median	Posterior SD	95% CS LL	95% CS UL

β_0	0.165	0.101	-0.039	0.369	-0.129	0.109	-0.345	0.082
β_1	-0.012	0.118	-0.252	0.227	-0.003	0.139	-0.274	0.257
τ_1^2 (Low)	0.034	0.036	0.000	0.096[3]	0.057	0.070	0.002	0.252
τ_2^2 (High)	0.103	0.052	0.020	0.159	0.110	0.079	0.025	0.325
ϕ	1.692	0.089	1.531	1.880[4]				
σ_ε^2					0.570	0.088	0.410	0.741
	Likelihood Ratio Test of $H_0: \tau_1^2 = \tau_2^2$				95% Credible Set for $\tau_1^2 - \tau_2^2$			
	-2 RQL log-like: Model 1	-2 RQL log-like: Model 2	LRT χ^2_1 statistic	Naïve *p*-value	Posterior Median	Posterior SD	95% CS LL	95% CS UL
	2475.64	2476.82	1.17	0.279	-0.046	0.107	-0.259	0.155

*NOTES: RPL = restricted pseudo-likelihood. LSE = Linearized Standard Error. CI = Confidence Interval. LL = Lower Limit. UL = Upper Limit. Log-like = log-likelihood. Model 1 = Model with unequal random effect variances in two groups. Model 2 = Model with random effect variances in two groups constrained to be equal.

**NOTES: CS = credible set. Inferences based on 1,074 effective simulated draws, from 5,000 draws with 2,500 draws discarded as burn-in. Gelman-Rubin (1992) $\hat{R}$ = 1 for three chains for each parameter. Posterior predictive p-value for simulated means = 0.169; posterior predictive p-value for simulated SDs = 0.335. DIC = 2190.1.

Both approaches generally agree in terms of inferences concerning the individual parameters in this model. The two experience groups do not differ in terms of mean measured parity, and the mean parity measured cannot be considered different from 1. In addition, regardless of how the overdispersion is parameterized, there is evidence of overdispersion in the parity data, providing support for the model being fitted. With regard to the question of whether the two variance components are significantly different, there does appear to be more variance in the higher experience group, but both the naïve likelihood ratio test and the 95% credible set for the difference in the two components suggest that the variance components are similar. There is not thus enough evidence to support the conclusion that these two groups of interviewers have different variance components. We do see that the 95% credible sets for the parameters in the model are wider than the approximate 95% confidence intervals from the frequentist approach, showing that additional sources of uncertainty in the parameter estimates are being accommodated by the Bayesian approach.

When fitting this model using the Bayesian approach and omitting the fixed effect of the low experience group (β_1), the posterior median of the difference in random effect variances between the groups was -0.049, and the 95% credible set for the difference based on draws from the posterior was (-0.265, 0.141), indicating that inference regarding this difference was not sensitive to the omission of the fixed group effect. This is not entirely surprising, given that this fixed group effect did not appear to be significantly different from zero based on the original analysis. Similar results were also found when omitting the fixed group effect in the frequentist analysis.

[3] The lower and upper 95% confidence limits computed for the variances of the random effects in the two groups in Table 2, Table 3, and Table 4 are *profile likelihood bounds* (SAS, 2010).

[4] The Wald-type 95% confidence limits for the overdispersion parameter ϕ (with a lower boundary constraint of zero) reported in Table 2, Table 3, and Table 4 are computed using a Satterthwaite approximation, as described in Milliken and Johnson (1992) and Burdick and Graybill (1992).

Table 3 compares results from the two approaches for the model comparing groups of interviewers defined by marital status (married vs. other).

Table 3: Comparisons of frequentist and Bayesian inferences for marital status of interviewer (married vs. other)

	Frequentist Analysis*				Bayesian Analysis**			
Parameter	RPL Estimate	LSE	95% CI LL	95% CI UL	Posterior Median	Posterior SD	95% CS LL	95% CS UL
β_0	0.141	0.080	-0.022	0.304	-0.137	0.089	-0.323	0.040
β_1	0.013	0.110	-0.211	0.237	-0.017	0.143	-0.296	0.257
τ_1^2 (Married)	0.126	0.060	0.026	0.190	0.151	0.092	0.042	0.381
τ_2^2 (Other)	0.003	0.024	0.000	0.053	0.023	0.040	<0.001	0.143
ϕ	1.684	0.088	1.524	1.871				
σ_ε^2					0.565	0.086	0.415	0.760
	Likelihood Ratio Test of $H_0: \tau_1^2 = \tau_2^2$				95% Credible Set for $\tau_1^2 - \tau_2^2$			
	-2 RQL log-like: Model 1	-2 RQL log-like: Model 2	LRT χ^2_1 statistic	Naïve *p*-value	Posterior Median	Posterior SD	95% CS LL	95% CS UL
	2473.12	2477.28	4.17	0.041	0.119	0.102	-0.029	0.360

*NOTES: RPL = restricted pseudo-likelihood. LSE = Linearized Standard Error. CI = Confidence Interval. LL = Lower Limit. UL = Upper Limit. Log-like = log-likelihood. Model 1 = Model with unequal random effect variances in two groups. Model 2 = Model with random effect variances in two groups constrained to be equal.

**NOTES: CS = credible set. Inferences based on 1,074 effective simulated draws, from 5,000 draws with 2,500 draws discarded as burn-in. Gelman-Rubin (1992) $\hat{R}$ = 1 for three chains for each parameter. Posterior predictive p-value for simulated means = 0.161; posterior predictive p-value for simulated SDs = 0.358. DIC = 2185.2.

We once again find no support for a fixed effect of interviewer marital status on mean parity, which is reassuring, and no evidence against the mean parity being equal to 1. The two approaches do, however, present evidence of larger variance among interviewers in the married group of interviewers. The results from the Bayesian approach once again show that the 95% credible sets for the parameters are wider, incorporating the additional uncertainty in the estimates that is not being accounted for by the frequentist approach. Both approaches once again provide support for overdispersion in the parity distribution, and the DIC for this model (2185.2) suggests a better fit than the model based on interviewer experience (DIC = 2190.1).

Importantly, the naïve likelihood ratio test following the frequentist approach would suggest that the two groups of interviewers have significantly different variance components (p = 0.041). In contrast, the 95% credible set for the difference in the two variance components based on the Bayesian approach does include 0, failing to provide support for a difference in variance components between these two groups of interviewers. The additional sources of uncertainty in estimating these variance components and small sample sizes that were not being accounted for in the asymptotic frequentist approach appear to be leading to a different conclusion in this case. It is worth nothing that a 90% credible set for the difference between these variance components is (-0.001, 0.317), providing weak support for a difference between the groups. Reasons for the

higher variance in the group of married interviewers should probably be investigated, given the weak evidence suggesting a possible difference in variance components between the groups.

When fitting the model using the Bayesian approach and omitting the fixed effect of the married group (β_1), the posterior median of the difference in random effect variances between the groups was 0.112, and the 95% credible set for the difference based on draws from the posterior was (-0.023, 0.335), indicating that inference regarding this difference was once again not sensitive to the omission of the fixed group effect. The frequentist model failed to converge when omitting the fixed marital group effect, indicating issues with estimation that were not present with the Bayesian approach.

Finally, Table 4 presents results from applying the two approaches to the model comparing the two groups of interviewers defined by other employment (yes vs. no).

Table 4: Comparisons of frequentist and Bayesian inferences for other employment status of interviewer

	Frequentist Analysis*				Bayesian Analysis**			
Parameter	RPL Estimate	LSE	95% CI LL	95% CI UL	Posterior Median	Posterior SD	95% CS LL	95% CS UL
β_0	0.214	0.102	0.007	0.420	-0.072	0.107	-0.282	0.130
β_1	-0.109	0.122	-0.357	0.140	-0.147	0.144	-0.422	0.132
τ_1^2 (Yes)	0.072	0.045	0.007	0.131	0.070	0.069	< 0.001	0.245
τ_2^2 (No)	0.071	0.046	0.007	0.134	0.093	0.075	0.009	0.302
ϕ	1.700	0.089	1.538	1.888				
σ_ε^2					0.585	0.093	0.410	0.780
	Likelihood Ratio Test of $H_0: \tau_1^2 = \tau_2^2$				95% Credible Set for $\tau_1^2 - \tau_2^2$			
	-2 RQL log-like: Model 1	-2 RQL log-like: Model 2	LRT χ^2_1 statistic	Naïve *p*-value	Posterior Median	Posterior SD	95% CS LL	95% CS UL
	2479.92	2479.92	0.00	0.994	-0.026	0.102	-0.247	0.181

*NOTES: RPL = restricted pseudo-likelihood. LSE = Linearized Standard Error. CI = Confidence Interval. LL = Lower Limit. UL = Upper Limit. Log-like = log-likelihood. Model 1 = Model with unequal random effect variances in two groups. Model 2 = Model with random effect variances in two groups constrained to be equal.

**NOTES: CS = credible set. Inferences based on 1,074 effective simulated draws, from 5,000 draws with 2,500 draws discarded as burn-in. Gelman-Rubin (1992) $\hat{R}$ = 1 for three chains for each parameter. Posterior predictive p-value for simulated means = 0.148; posterior predictive p-value for simulated SDs = 0.374. DIC = 2186.5.

The results in Table 4 provide little evidence of a difference in parity means or variance components between groups of interviewers defined by other employment status, whether following the frequentist or Bayesian approach. Inference regarding this difference was once again not sensitive to the omission of the fixed group effect, and similar results were found in the frequentist approach. In general, future studies should consider sensitivity of inferences to omission of these effects if different groups of interviewers do in fact have different expected values on the survey response of interest.

3.3 Checking Model Fit

The Bayesian approach enables one to check the fits of these models using posterior predictive checks. Posterior predictive p-values for the mean and standard deviation of parity presented in the notes for Table 2, Table 3, and Table 4 all suggest that the mean and standard deviation of parity are being fit well by the three models. Figure 2 presents a visual examination of the posterior predictive checks considered for each of the three models fitted using the Bayesian approach. In this figure, the observed mean or standard deviation is shown as a boldfaced vertical line, and the histogram shows the distribution of simulated replications of the mean and standard deviation based on the posterior draws of the parameters from each fitted model.

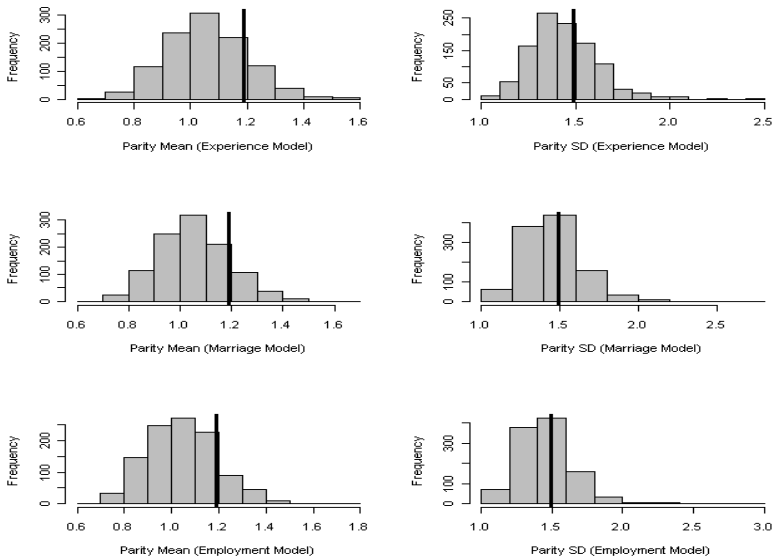

Figure 2: Posterior predictive checks for means and standard deviations of parity, based on the three models fitted using the Bayesian approach. Solid vertical lines represent means and SDs of the observed parity data, and histograms represent distributions of simulated means and standard deviations of parity values based on the fitted models

The standard deviation in particular is being fit well by the three models, and the observed mean seems entirely reasonable based on the fitted models.

4. Discussion

This paper has demonstrated Bayesian approaches to fitting HGLMs with heterogeneous random effect variance parameters in two independent groups of clusters and making inferences about differences in those variance parameters. Analyses of real survey data from the NSFG have shown how the Bayesian approaches do a better job than frequentist approaches of accommodating uncertainty in the estimation of parameters in these models, and lead to more appropriate inferences when the number of clusters under study is fairly small. Specifically, inferences when following the Bayesian approach to analyzing this problem can be based on 95% credible sets for the difference in the two variance components, defined by the differences in simulated draws of the two variance components from the joint posterior distribution for a given model. This approach provides a more natural form of inference for this problem than the more problematic likelihood ratio testing in the frequentist setting, which relies on asymptotic theory and should not be applied when using pseudo-likelihood estimation approaches. Software code

for replicating these analyses more generally in SAS, R, and BUGS can be found in the Appendix.

An application of the Bayesian approach to the problem of interviewer variance in survey methodology found weak support for higher interviewer variance in the measurement of parity among married NSFG interviewers. Although many explanations for this finding are certainly possible (e.g., married female interviewers varying in their ability to convey the true meaning of parity for some reason) and additional analyses should consider sensitivity of these specific results to outliers in the measurement of parity, this technique would probably be more applicable in practice when a randomized experiment is designed to assess whether a particular training technique serves to reduce interviewer variance. Interest in this case would lie in making inferences about the difference in variance components for particular survey measures between interviewers receiving a specialized form of training and interviewers receiving standard training.

Extensions of the Bayesian approach considered in this specific application are certainly possible. HGLMs for other types of non-normal variables (binary survey variables, nominal categorical survey variables, etc.) and HLMs for normal variables could be fitted using the same approaches presented in this paper. The approach could also be extended to comparisons of the variance components for more than two groups of clusters (or interviewers), and Bayesian methods could accommodate the possibility of measurement error in the covariates used to define the groups of clusters. Simulations of posterior differences between any two variance components could be computed as long as draws of the variance components in the various groups from the joint posterior distribution are feasible in the algorithm used by BUGS. The fairly simple models used in this paper also considered a single fixed group effect only. Investigations seeking to explain significant interviewer variance in a particular group (e.g., among married interviewers) would more than likely add additional interviewer-level covariates to these models, and that extension would also be straightforward.

This paper also did not consider another rich aspect of the Bayesian approach, in that posterior draws of the 38 random interviewer effects in the models were also generated by the BUGS Gibbs sampling algorithm. These draws would enable survey managers to make inferences about the effects specific interviewers are having on particular survey measures. Consistent and regular updating of these posterior distributions as data collection progresses would enable survey managers to intervene when the posterior distributions for particular interviewers suggest that these interviewers are having non-zero effects on the survey measures.

References

Biemer, P.P., and Trewin, D. (1997). Chapter 27: A Review of Measurement Error Effects on the Analysis of Survey Data. *Survey Measurement and Process Quality*. Editors Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin. Wiley-Interscience. Pp. 603-632.

Burdick, R.K. and Graybill, F.A. (1992), *Confidence Intervals on Variance Components,* New York: Marcel Dekker.

Carlin, B.P., and Louis, T.A. (2009). *Bayesian Methods for Data Analysis*. Chapman and Hall / CRC Press.

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall / CRC Press.

Fowler, F.J., and Mangione, T.W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Anaylsis*. Chapman and Hall / CRC Press.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Gilks, W.R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

Groves, R.M. (2004). Chapter 8: The Interviewer as a Source of Survey Measurement Error. *Survey Errors and Survey Costs (2nd Edition)*. Wiley-Interscience.

Groves, R.M., Mosher, W.D., Lepkowski, J.M. and Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. National Center for Health Care Statistics. *Vital Health Statistics*, 1(48).

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association*, 57, 92-115.

Mahalanobis, P.C. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Milliken, G.A. and Johnson, D.E. (1992), *Analysis of Messy Data, Volume 1: Designed Experiments*, New York: Chapman & Hall.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, Berlin.

O'Muircheartaigh, C., and Campanelli, P. (1998). The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. *Journal of the Royal Statistical Society, Series A*, 161 (1), 63-77.

Pinheiro, J.C. and Chao, E.C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 15, 58–81.

SAS Institute, Inc. (2010). Online Documentation for the GLIMMIX Procedure.

Schober, M., and Conrad, F. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? *Public Opinion Quarterly*, 61, 576-602.

Sinha, S.K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37(2), 219-234.

West, B.T. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Submitted to Public Opinion Quarterly and Under Revision, April 2010.*

Zhang, D. and Lin, X. (2010). Variance component testing in generalized linear mixed models for longitudinal / clustered data and other related topics. *Random Effect and Latent Variable Model Selection*. Springer Lecture Notes in Statistics, Volume 192.