Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

# Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers


by Brady T. West and Michael R. Elliott

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**email** at infostats@statcan.gc.ca,

**telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

**Depository Services Program**

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada



Cette publication est aussi disponible en français.

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

## Standard symbols

The following symbols are used in Statistics Canada publications:

| | |
|---|---|
| . | not available for any reference period |
| .. | not available for a specific reference period |
| ... | not applicable |
| 0 | true zero or a value rounded to zero |
| $0^s$ | value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded |
| $^p$ | preliminary |
| $^r$ | revised |
| x | suppressed to meet the confidentiality requirements of the *Statistics Act* |
| $^E$ | use with caution |
| F | too unreliable to be published |
| * | significantly different from reference category ($p < 0.05$) |

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 163-188
Statistics Canada, Catalogue No. 12-001-X


# Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers

**Brady T. West and Michael R. Elliott**[1]

Abstract

Survey methodologists have long studied the effects of interviewers on the variance of survey estimates. Statistical models including random interviewer effects are often fitted in such investigations, and research interest lies in the magnitude of the interviewer variance component. One question that might arise in a methodological investigation is whether or not different groups of interviewers (e.g., those with prior experience on a given survey vs. new hires, or CAPI interviewers vs. CATI interviewers) have significantly different variance components in these models. Significant differences may indicate a need for additional training in particular subgroups, or sub-optimal properties of different modes or interviewing styles for particular survey items (in terms of the overall mean squared error of survey estimates). Survey researchers seeking answers to these types of questions have different statistical tools available to them. This paper aims to provide an overview of alternative frequentist and Bayesian approaches to the comparison of variance components in different groups of survey interviewers, using a hierarchical generalized linear modeling framework that accommodates a variety of different types of survey variables. We first consider the benefits and limitations of each approach, contrasting the methods used for estimation and inference. We next present a simulation study, empirically evaluating the ability of each approach to efficiently estimate differences in variance components. We then apply the two approaches to an analysis of real survey data collected in the U.S. National Survey of Family Growth (NSFG). We conclude that the two approaches tend to result in very similar inferences, and we provide suggestions for practice given some of the subtle differences observed.

**Key Words:** Interviewer variance; Bayesian analysis; Hierarchical generalized linear models; Likelihood ratio testing.

## 1 Introduction

Between-interviewer variance in survey methodology (e.g., West, Kreuter and Jaenichen 2013; West and Olson 2010; Gabler and Lahiri 2009; O'Muircheartaigh and Campanelli 1998; Biemer and Trewin 1997; Kish 1962) occurs when survey responses nested within interviewers are more similar than responses collected from different interviewers. Between-interviewer variance can increase the variance of survey estimates of means, and may arise due to correlated response deviations introduced by an interviewer (e.g., Biemer and Trewin 1997), given the complexity of survey questions (e.g., Collins and Butcher 1982) or interactions between the interviewer and the respondent (e.g., Mangione, Fowler and Louis 1992), or nonresponse error variance among interviewers (West et al. 2013; Lynn, Kaminska and Goldstein 2011; West and Olson 2010).

Survey research organizations train interviewers to eliminate this component of variance in survey estimates, as it is sometimes larger than the component of variance due to cluster sampling (Schnell and Kreuter 2005). In reality, an interviewer variance component can never be equal to 0 (which would imply that means on the variable of interest are identical across interviewers), but survey managers aim to minimize this component via specialized interviewer training. For example, interviewers may practice the

1. Brady T. West, Survey Methodology Program, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48106. E-mail: bwest@umich.edu; Michael R. Elliott, Survey Methodology Program, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48106. E-mail: mrelliot@umich.edu.

administration of selected questions under the direct supervision of training staff, and then receive feedback on any variance in administration that is noted by the staff (in an effort to standardize the administration; see Fowler and Mangione 1990). In some non-interpenetrated designs, where interviewers are generally assigned to work exclusively in a single primary sampling area (e.g., the U.S. National Survey of Family Growth; see Lepkowski, Mosher, Davis, Groves and Van Hoewyk 2010), interviewer effects and area effects are confounded, preventing estimation of the variance in survey estimates that is uniquely attributable to the interviewers. Elegant interpenetrated sample designs (Mahalanobis 1946) enable interviewers to work in multiple sampling areas, and in these cases, cross-classified multilevel models can be used to estimate the components of variance due to interviewers and areas (e.g., Durrant, Groves, Staetsky and Steele 2010; Gabler and Lahiri 2009; Schnell and Kreuter 2005; O'Muircheartaigh and Campanelli 1999; O'Muircheartaigh and Campanelli 1998).

In general, estimating the overall magnitude of interviewer variance in the measures of a given survey variable or data collection process outcome is a useful exercise for survey practitioners. If random subsamples of sample units are assigned to interviewers following an interpenetrated design, one can estimate the component of variance due to interviewers and subsequently the unique effects of interviewers on the variance of an estimated survey mean (e.g., Groves 2004, p. 364). Large estimates can indicate potential measurement difficulties that certain interviewers are experiencing, or possible differential success in recruiting particular types of sampled units. Given a relatively large estimate of an interviewer variance component and an appropriate statistical test indicating that the component is significantly larger than zero (or "non-negligible", given that variance components technically cannot be exactly equal to zero; see Zhang and Lin 2010), survey managers can use various methods to compute predictions of the random effects associated with individual interviewers, and identify interviewers who may be struggling with particular aspects of the data collection process.

While the estimation of interviewer variance components and subsequent adjustments to interviewer training and data collection protocols have a long history in the survey methodology literature (see Schaeffer, Dykema and Maynard 2010 for a recent review), no studies in survey methodology to date have examined the alternative approaches that are available to survey researchers for *comparing* variance components in two independent groups of survey interviewers. In general, alternative statistical approaches are available for estimating interviewer variance components, and estimates (and corresponding inferences about the variance components) may be sensitive to the estimation methods that a survey researcher employs. The same is true for survey researchers who may desire to compare the variance components associated with different groups of interviewers, for various reasons (e.g., identifying groups that need more training or more optimal modes for certain types of questions): different statistical approaches to performing these kinds of comparisons exist, and inferences about the differences may be sensitive to the approach used. With this paper, we aim to evaluate alternative frequentist and Bayesian approaches to making inference about the differences in variance components between two independent groups of survey interviewers, and provide practical guidance to survey researchers interested in this type of analysis.

The paper is structured as follows. In Section 2, we introduce the general modeling framework that enables these comparisons of interviewer variance components for both normal and non-normal (e.g., binary, count) survey variables, and review existing literature comparing the frequentist and Bayesian approaches to estimation and inference, highlighting the advantages and disadvantages of each approach.

We then present a simulation study in Section 3, evaluating the ability of the two approaches to efficiently estimate differences in variance components between two hypothetical groups of interviewers. Section 4 applies the two approaches to real survey data collected in the U.S. National Survey of Family Growth (NSFG) (Lepkowski et al. 2010; Groves, Mosher, Lepkowski and Kirgis 2009). Finally, Section 5 offers concluding thoughts, suggestions for practitioners, and directions for future research. We include SAS, R, and WinBUGS code that readers can use to implement the two approaches in the Appendix.

# 2 Alternative approaches for comparing variance components in Hierarchical Generalized Linear Models

We first consider a general class of models that survey researchers can employ to compare variance components in different groups of interviewers. Hierarchical Generalized Linear Models (HGLMs) are flexible analytic tools that can be used to model observations on both normal and non-normal (e.g., binary, count) survey variables of interest, where observations nested within the same interviewer cannot be considered independent (Raudenbush and Bryk 2002; Goldstein 1995). We consider alternative approaches to making inferences about interviewer variance components in a specific class of HGLMs, where the interviewer variance components for two independent groups of interviewers defined by a known interviewer characteristic need not be equal. This type of HGLM can be written as

$$\begin{aligned} & g\left(E\left[y_{ij} \mid u_i\right]\right) = \beta_0 + \beta_1 I\left(Group = 1\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i \\ & u_{i(1)} \sim N\left(0, \tau_1^2\right),\ u_{i(2)} \sim N\left(0, \tau_2^2\right), \end{aligned} \tag{2.1}$$

where $g\left(x\right)$ is the link function relating a transformation of the expected value of the dependent variable, $y_{ij}$, to the linear combination of the fixed and random effects (e.g., $g\left(x\right) = log\left[x/\left(1-x\right)\right]$ for an assumed Bernoulli distribution [binary outcome], $g\left(x\right) = log\left(x\right)$ for an assumed Poisson distribution [count outcome]), $i$ is an index for the interviewer, $j$ is an index for the respondent nested within an interviewer, and $I\left(\bullet\right)$ represents an indicator variable, equal to 1 if the condition inside the parentheses is true and 0 otherwise. The random interviewer effects from Group 1, $u_{i(1)}$, are assumed to follow a normal distribution with mean 0 and variance $\tau_1^2$, while the random interviewer effects from Group 2, $u_{i(2)}$, are assumed to follow a normal distribution with mean 0 and variance $\tau_2^2$. Other distributions may be posited for the random effects, and the general model in (2.1) can accommodate over-dispersion in the observed dependent variable relative to the posited distribution for that variable. The key aspect of the specification in (2.1) is that random effects for different groups of interviewers have *different* variances. The fixed effect parameter $\beta_1$ in (2.1) represents a fixed effect of Group 1 on the outcome relative to Group 2 in the HGLM, and fixed effects of other covariates can easily be included. Similarly, additional subgroups of interviewers can be considered by including additional random effects $u_{i(k)}$, for $k > 2$. Analytic interest lies in the magnitude of the difference in the variance components.

Models of the form in (2.1) can be applied when methodological studies are designed to compare two different groups of interviewers in terms of their variance components. For example, there exists a debate

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 (2.1), including indicator variables for the two groups of interviewers. This same approach could be used to compare the interviewer variance components in two groups of interviewers randomly assigned to different data collection modes (e.g., CAPI vs. CATI). To date, no published studies have attempted these kinds of comparisons, but they are important for understanding the overall impacts of these design decisions on the mean squared error (MSE) of survey estimates.

Frequentist approaches to the estimation of parameters in HGLMs rely on various numerical or theoretical approaches to approximating complicated likelihood functions, especially for models such as (2.1) that involve complex random effects structures (e.g., Faraway 2006, Chapter 10; Molenberghs and Verbeke 2005). In general, inferences are based on these approximate likelihood-based approaches, which include residual pseudo-likelihood (which is different from the pseudo-maximum likelihood estimation approach developed by Binder (1983) for design-based analyses of data from complex sample surveys), penalized quasi-likelihood, and maximum likelihood based on a Laplace approximation. Previous work has found favorable simulation results for the residual pseudo-likelihood approach, which indicate nearly unbiased estimation of the variance components in an HGLM as compared to maximum likelihood using Laplace approximation or adaptive quadrature (Pinheiro and Chao 2006). These findings are similar to the case of restricted maximum likelihood (REML) estimation in a model for a normally distributed outcome variable. For binary outcome variables, marginal or penalized quasi-likelihood techniques can lead to downward bias in parameter estimates and convergence problems, and fully Bayesian approaches may have favorable properties in this case (Browne and Draper 2006; Rodriguez and Goldman 2001). We therefore consider the residual pseudo-likelihood approach in the simulations and applications presented in this study, and contrast this approach with a fully Bayesian approach.

There are two approaches available for making inference about differences in variance components in the frequentist setting. The first approach involves 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. 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 test statistic is computed by fitting a constrained version of the model in (2.1), with the random effect variance components in the two groups specified as equal, and then fitting the model with the more general form in (2.1). The positive difference in the approximate -2 log-likelihood values of these two models is then computed, and referred to a chi-square distribution with one degree of freedom.

The second approach involves computing the difference of the pseudo-ML estimates, $\hat{\tau}_1 - \hat{\tau}_2$, and an associated 95% Wald-type confidence interval for the difference, given by $\hat{\tau}_1 - \hat{\tau}_2 \pm 1.96\sqrt{\hat{\text{var}}(\hat{\tau}_1) + \hat{\text{var}}(\hat{\tau}_2) - 2\hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)}$. This interval requires asymptotic estimates of the variances and covariances of the two estimated variance components, which are computed based on the Hessian (second derivative) matrix of the objective function used for the maximum likelihood estimation procedure. If the resulting Wald interval includes zero, one would conclude that there is not enough evidence against the null hypothesis. Confidence intervals for differences in variance components can also be computed using inversions of profile likelihood tests (e.g., Viechtbauer 2007), although standard software does not include options for implementing this procedure (to our knowledge).

These two frequentist approaches to making inference about differences in interviewer variance components do have limitations. When the number of interviewers in each group is small (say, less than 30; see Hox (1998) for discussion), asymptotic results for the likelihood ratio test (Zhang and Lin 2010) may no longer hold. Frequentist (maximum likelihood) methods also tend to overstate the precision of estimates, given that they ignore the uncertainty in estimates of the variance components (Carlin and Louis 2009, p. 335-336), which is especially problematic for small samples (Goldstein 1995, p. 23). Bayesian approaches allow analysts to place prior distributions on variance components to reflect this uncertainty, unlike frequentist approaches. Furthermore, Molenberghs and Verbeke (2005, p. 277) argue that likelihood ratio tests should not be used to test hypotheses when models are fitted using pseudo-likelihood methods. Approximate maximum likelihood estimation methods can also lead to invalid (i.e., negative) estimates of variance components in these models when variance components are very small. Software that does not use estimation procedures constraining these variance components to be greater than zero generally responds to this problem by setting negative estimates of variance components equal to zero (with no accompanying standard error), which prevents computation of the Wald-type confidence interval described above.

A Bayesian approach to fitting the HGLMs described in (2.1) uses the MCMC-based Gibbs sampler and the adaptive rejection sampling methodology (Gilks and Wild 1992) to simulate draws from the posterior distribution for the parameters in the model defined in (2.1). In general, the posterior distributions for the parameters in an HGLM are not of known distributional forms and need to be simulated (Gelman, Carlin, Stern and Rubin 2004, Section 16.4). Diffuse, non-informative priors for the fixed effects and the variance components in (2.1) can be specified for the simulations, to let the data provide the most information about the posterior distributions of the parameters (Gelman and Hill 2007; Gelman 2006, Section 7). This approach enables inferences based on simulated draws from the marginal posterior distributions of the two fixed effect parameters, the two variance parameters, the 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 interviewer-level characteristic, computed using the simulated draws of the two variance components.

Given that traditional hypothesis tests are not meaningful in the Bayesian setting, Bayesian inference will focus on the difference in the interviewer variance components. Inference for the difference is 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 $\tau_1^{2(d)} - \tau_2^{2(d)}$, can 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 a given analysis, several thousand draws from the posterior distribution can be generated using the Gibbs sampler, with a large number of initial draws discarded as burn-in draws, and the effective number of simulation draws will be computed based on the number of burn-in draws (Gelman and Hill 2007, Chapter 16). If the resulting 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 can 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 fixed effects 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) and computing a 95% credible set for that parameter. While such estimates can also be obtained in the frequentist approach, as noted previously, the Bayesian approach does not require asymptotic assumptions and incorporates the variability in the estimated variance components into the computation of the 95% credible sets via the simulated draws.

Multiple (typically three) Markov chains can be run in parallel in the iterative Gibbs sampling algorithm to simulate random walks through the space of the joint posterior distribution. The Gelman-Rubin $\hat{R}$ statistic, representing (approximately) the square root of the variance of the mixture of the chains divided by the average within-chain variance (Gelman and Rubin 1992), can be used to assess convergence (or mixing) of the chains for each parameter. Values less than 1.1 on this statistic can be considered as evident of convergence of the chains for a given parameter. Posterior draws of the parameters can be pooled from the three chains to generate the final effective sample size of draws used for inferences.

The Bayesian approach outlined above is also not without limitations. The selection of the prior distributions used to compute the posterior distribution for the parameters in (2.1) is essentially arbitrary, and depends on the choices of a given analyst and the amount of prior information available. Furthermore, the choice of the prior distribution can become crucial when there is a small number of interviewers (say, less than 20), where different priors can lead to very different inferences regarding the variance components (Lambert, Sutton, Burton, Abrams and Jones 2005); the use of prior information about the variance components can increase efficiency relative to the use of non-informative priors in these cases. Model misspecification is also a distinct possibility depending on the survey variable being modeled, which is also a limitation of the frequentist approach. Computational demand may also be an issue with the Bayesian (Gibbs sampling) approach (Browne and Draper 2006), especially if one desires comparisons of interviewer variance components for a large number of survey variables (with potentially different distributions) and there are a relatively large number of interviewers; this may not be as problematic with recent advances in hardware speed and algorithm efficiency. Finally, analysts may not be comfortable

with the available software for Bayesian approaches, so there may be a learning curve associated with implementation of this approach.

Several previous articles have compared these alternative frequentist and Bayesian approaches using simulation studies. Chaloner (1987) considered one-way ANOVA models with random effects for unbalanced data (similar to the case in this study, where interviewers have different workloads), and found lower empirical MSE values for posterior modes of the variance components when following the Bayesian approach and using non-informative priors than for the frequentist (maximum likelihood) approach. Van Tassell and Van Vleck (1996) reported that the Gibbs sampler (using either informative or non-informative prior distributions) and REML both produce empirically unbiased estimates of variance components that tend to be extremely similar. Browne and Draper (2006) also found that both approaches can lead to unbiased estimates, with the more "automatic" nature of frequentist approaches being an attractive feature. In the context of predicting means for small areas using models with random area effects, Singh, Stukel and Pfeffermann (1998) reported that Bayesian MSE approximations for the predictions have good frequentist properties, but that the Bayesian method tends to produce larger frequentist biases and prediction MSEs than frequentist methods. Farrell (2000) found that the Bayesian approach resulted in slightly more accurate predictions of small area proportions, with little differences in coverage rates or bias between the two approaches. Ugarte, Goicoa and Militino (2009) also found that the two approaches performed quite similarly in an application involving the detection of high-risk areas for disease. These authors point out that the relative computational simplicity of the frequentist approach is attractive in light of these findings. In general, based on the literature in this area, we anticipate similar performance of the two methods in the case of comparing interviewer variance components, and we evaluate this expectation using a simulation study (Section 3).

While there exist many software procedures for fitting multilevel models and estimating variance components using both frequentist and Bayesian methods (see West and Galecki 2011 for a review), the frequentist approach to the specific comparison of variance components discussed in this paper is only readily implemented in the GLIMMIX procedure of SAS/STAT (SAS 2010), through the COVTEST statement with the HOMOGENEITY option (which assumes that a GROUP variable has been specified in the RANDOM statement, indicating different groups of clusters with random effects arising from different distributions). We are not aware of any other procedures that readily implement the frequentist comparison approach at the time of this writing. Example code that can be used for fitting these models using the GLIMMIX procedure is available in the Appendix. The Bayesian approach to comparing the variance components can be implemented in the BUGS (Bayesian Inference using Gibbs Sampling) software (see References Section for more details). We also include example code that implements this approach by calling WinBUGS from R in the Appendix.

# 3 Simulation Study

We conducted a small simulation study to examine the empirical properties of these two alternative approaches. Data on two hypothetical survey variables of interest (one normally distributed, one Bernoulli distributed) were simulated according to the following two super-population models:

$$y_{ij} = 45 + 5 \times I\left(Group = 2\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i + \varepsilon_{ij}$$
$$u_{i(1)} \sim N\left(0,1\right),\ u_{i(2)} \sim N\left(0,2\right),\ \varepsilon_{ij} \sim N\left(0,64\right) \tag{3.1}$$

$$P\left(y_{ij} = 1\right) = \frac{\exp\left[-1 + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i\right]}{1 + \exp\left[-1 + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i\right]} \tag{3.2}$$
$$u_{i(1)} \sim N\left(0,0.03\right),\ u_{i(2)} \sim N\left(0,0.13\right).$$

The notation used here is consistent with that used in (2.1). Values on the second Bernoulli variable were generated for hypothetical cases according to the logistic regression model specified in (3.2). To obtain the observed Bernoulli variable, a random draw was obtained from a UNIFORM(0,1) distribution, and the variable was set to 1 if the random draw was less than or equal to the predicted probability, and 0 otherwise. For one hypothetical group of interviewers at a time, random interviewer effects were drawn, and values for cases within each interviewer were then generated according to the specified model.

We generated 200 samples of hypothetical cases and simulated data for each variable, with 50 hypothetical interviewers in one group collecting data from 50 hypothetical cases each ($n = 2,500$ for each group of interviewers). We then generated an additional 200 samples in a small-sample scenario, with 20 interviewers in each group collecting data from ten hypothetical cases each ($n = 200$ for each group of interviewers). The choices of the variance components in (3.1) correspond to intra-interviewer correlations of 0.015 and 0.030 for the two hypothetical groups of interviewers, while the choices of the variance components in (3.2) correspond to intra-interviewer correlations of 0.009 and 0.038. All of these values would be considered plausible in a face-to-face or telephone survey setting (West and Olson 2010). The known differences in variance components between the groups are therefore 1 for the normal variable, and 0.1 for the Bernoulli variable.

Given these known values for the interviewer variance components in the hypothetical population, we applied each method described in Section 2 [using diffuse, non-informative, uniform priors for the variance components, per recommendations of Gelman (2006, Section 7)] to each hypothetical sample. We computed the following empirical measures for comparison purposes: 1) the empirical and relative bias of the estimator; 2) the empirical MSE of the estimator; 3) the "frequentist" coverage of the 95% Wald-type intervals (when using the frequentist approach) and the 95% credible sets (when using the Bayesian approach); and 4) the average widths of the 95% Wald-type intervals and the credible sets. The number of Wald-type intervals that could not be computed due to estimated variance components of 0 (with no accompanying standard errors) was also recorded in each case. All simulations were performed using SAS, R, and BUGS, and simulation code is available upon request.

Table 3.1 presents the results of the simulation study. The results suggest that for moderate-to-large samples of interviewers and respondents, both approaches yield estimators of the difference in variance components that have fairly small bias, as anticipated. The frequentist approach was found to yield estimators with smaller empirical MSE values; this is not entirely surprising, given the additional variability in the Bayesian estimates introduced by accounting for uncertainty in the prior distributions of the parameters with non-informative priors. The use of more informative priors may improve the efficiency of the Bayesian estimates. In the large sample setting, the 95% confidence intervals and credible sets computed for the difference in variance components appear to have acceptable coverage properties, with the Bayesian approach having slight under-coverage.

**Table 3.1**
**Results of simulation study comparing the empirical properties of the frequentist and Bayesian approaches to making inference about the differences in interviewer variance components.**

| Sample Sizes | | Frequentist Approach | Bayesian Approach |
|---|---|---|---|
| | **Normal Y** | | |
| | Empirical Bias | -0.0498 | -0.0189 |
| | Relative Bias | -4.98% | -1.89% |
| | Empirical MSE | 0.6546 | 0.8134 |
| | 95% CI/CS Coverage | 0.960 | 0.920 |
| | Mean 95% CI/CS Width | 3.1689 | 3.6283 |
| 50 interviewers / group | % of Wald CIs Invalid | 0.0% | -- |
| 50 cases / interviewer | **Bernoulli Y** | | |
| ( $n = 2,500$ / group) | Empirical Bias | -0.0020 | -0.0046 |
| | Relative Bias | -2.0% | -4.6% |
| | Empirical MSE | 0.0029 | 0.0033 |
| | 95% CI/CS Coverage | 0.938 | 0.940 |
| | Mean 95% CI/CS Width | 0.2142 | 0.2372 |
| | % of Wald CIs Invalid | 11.5% | -- |
| | **Normal Y** | | |
| | Empirical Bias | -0.2341 | -0.3508 |
| | Relative Bias | -23.41% | -35.08% |
| | Empirical MSE | 6.9873 | 6.2869 |
| | 95% CI/CS Coverage | 1.000 | 0.995 |
| | Mean 95% CI/CS Width | 16.6313 | 18.3574 |
| 20 interviewers / group | % of Wald CIs Invalid | 54.0% | -- |
| 10 cases / interviewer | **Bernoulli Y** | | |
| ( $n = 200$ / group) | Empirical Bias | -0.0348 | -0.0196 |
| | Relative Bias | -34.8% | -19.6% |
| | Empirical MSE | 0.0345 | 0.0861 |
| | 95% CI/CS Coverage | 1.000 | 0.980 |
| | Mean 95% CI/CS Width | 1.2604 | 1.7970 |
| | % of Wald CIs Invalid | 65.5% | -- |

Notably, 11.5% of the 95% Wald-type confidence intervals could not be computed when analyzing the binary outcome for the larger samples, due to one of the estimated variance components being equal to zero (with no standard error). This "failure" rate for the Wald intervals became much worse for both variables in the smaller samples, where both methods also produced inefficient estimates with a negative bias. The frequentist approach can therefore provide an estimate of the difference and associated confidence intervals that work well in larger samples with normally distributed variables, but in small samples or even moderate-to-large samples with non-normal variables, the simple Wald-type intervals that can be computed using standard software may fail a fairly substantial fraction of the time. This is due to

the fact that the Hessian matrix is not invertible when an estimated variance component is set to zero (i.e., the likelihood can't be approximated by a quadratic). Collectively, these simulation results therefore suggest that: 1) both approaches will perform similarly well when applied to real survey data with moderate-to-large samples of interviewers and respondents; 2) the Bayesian approach may be the better option if intervals (or credible sets) for the difference are desired; and 3) caution is advised when applying either method to relatively small samples of interviewers and respondents.

# 4 Application: The U.S. National Survey of Family Growth (NSFG)

We now apply the frequentist and Bayesian approaches to real survey data collected in the seventh cycle of the NSFG (June 2006 – June 2010). The original design of this cycle of the NSFG (Groves et al. 2009) called for 16 quarters of data collection from a continuous sample that was nationally representative when it was completed in June 2010. The data analyzed in this paper were collected from a national sample of 11,609 females between the ages of 15 and 44, by 87 female interviewers (with varying sample sizes for each interviewer). For more details on the design and operation of the seventh cycle of the NSFG, see Lepkowski et al. (2010) or Groves et al. (2009).

Each of the 87 interviewers has information available on her age (47.1% are age 55 or greater), years of experience (43.7% have five or more years of experience), number of children (33.3% have two or more children), marital status (19.5% have never been married), other employment (46.0% have other jobs), college education (57.5% completed a four-year college degree), previous experience working on NSFG (82.8% have worked on previous cycles), and ethnicity (81.6% are white). These observable interviewer-level characteristics will be used to divide the interviewers into two groups (in the absence of an ideal randomized experiment, like that described in Section 2).

Each of the 11,609 female respondents has their parity (or count of live births) and an indicator of current sexual activity (indicated by at least one current male partner or at least one male partner in the past 12 months) measured and available for analysis. While these measures seem fairly simple, the concepts being measured may be communicated differently by different interviewers (resulting in interviewer variance). The primary analytic question is whether these different groups of female interviewers have significantly different variance components for these particular survey variables.

We first consider an HGLM for the parity variable. Let $Y$ be a Poisson random variable with parameter $\lambda$. We allow for overdispersion (or extra-Poisson dispersion) in $Y$, which is quite common in count variables (for example, the mean parity for the sample of 11,609 females is 1.19, and the variance of the measured parity values is 1.99). Following Hilbe (2007) and Durham, Pardoe and Vega (2004), we let $\lambda = r\mu$, where $r$ is a GAMMA$\left(\alpha^{-1}, \alpha^{-1}\right)$ random variable. It then follows that $Y$ has a negative binomial distribution with mean $\mu$ and scale parameter $\alpha$:

$$E(Y) = E(\lambda) = E(r\mu) = \mu E(r) = \mu$$

$$\text{var}(Y) = E(\lambda) + \text{var}(\lambda) = E(r\mu) + \text{var}(r\mu) = \mu E(r) + \mu^2 \text{var}(r) = \mu(1 + \alpha\mu)$$

We specify an HGLM for the observed value of parity on female respondent $j$ interviewed by interviewer $i$, $y_{ij}$, as follows:

$$
\begin{aligned}
&y_{ij} \sim Poisson\left(\lambda_i\right),\ \lambda_i = r_i\mu_i \\
&r_i \sim Gamma\left(\alpha^{-1},\alpha^{-1}\right) \\
&\log\left(\mu_i\right) = \beta_0 + \beta_1 I\left(Group=1\right)_i + u_{i(1)} I\left(Group=1\right)_i + u_{i(2)} I\left(Group=2\right)_i \\
&u_{i(1)} \sim N\left(0,\tau_1^2\right), \qquad u_{i(2)} \sim N\left(0,\tau_2^2\right).
\end{aligned}
\tag{4.1}
$$

In this multilevel negative binomial regression model, $\exp\left(\beta_0\right)$ represents the expected parity for Group 2, $\exp\left(\beta_1\right)$ 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, and $u_{i(2)}$ is a random effect associated with interviewer $i$ in Group 2.

Next, we consider an HGLM for the binary indicator of current sexual activity. Let $z_{ij}=1$ if a female respondent $j$ indicates current sexual activity to interviewer $i$, and 0 otherwise. We specify the following model for this binary indicator:

$$
\begin{aligned}
&z_{ij} \sim Bernoulli\left(p_i\right) \\
&\ln\left[p_i/\left(1-p_i\right)\right] = \beta_0 + \beta_1 I\left(Group=1\right)_i + u_{i(1)} I\left(Group=1\right)_i + u_{i(2)} I\left(Group=2\right)_i \\
&u_{i(1)} \sim N\left(0,\tau_1^2\right), \qquad u_{i(2)} \sim N\left(0,\tau_2^2\right).
\end{aligned}
\tag{4.2}
$$

In this model, $\exp\left(\beta_0\right)$ represents the expected odds of current sexual activity for Group 2, $\exp\left(\beta_1\right)$ represents the expected multiplicative change in the odds of current sexual activity for Group 1 relative to Group 2, $u_{i(1)}$ is a random effect associated with interviewer $i$ in Group 1, and $u_{i(2)}$ is a random effect associated with interviewer $i$ in Group 2.

We fit models (4.1) and (4.2) using the two approaches described in Section 2. For the frequentist approach, based on recommendations from the literature discussed in Section 2, we estimated the parameters in these models using residual pseudo-likelihood (RPL) estimation, as implemented in the GLIMMIX procedure in the SAS/STAT software. All frequentist analyses presented in this section were repeated using adaptive quadrature to approximate the likelihood functions, and the primary results did not change; in addition, the use of adaptive quadrature led to longer estimation times.

For the Bayesian approach, the following non-informative prior distributions for these parameters were used. These prior distributions were 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 non-informative prior distributions for variance parameters in hierarchical models with a reasonably large number (i.e., more than five) of groups (or interviewers, in the present context):

$$
\begin{aligned}
&\beta_0 \sim N\left(0,100\right) \qquad \beta_1 \sim N\left(0,100\right) \\
&\tau_1^2 \sim Uniform\left(0,10\right) \qquad \tau_2^2 \sim Uniform\left(0,10\right) \\
&\ln(\alpha) \sim N\left(0,100\right).
\end{aligned}
$$

The non-informative priors for the fixed effects and the (natural log transformed) scale parameter for the negative binomial count variable (parity) indicate an expectation that these parameters 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 (untransformed) scale parameter and variance components ranging between 0 and 5, these priors are all fairly diffuse, expressing little prior knowledge about these parameters and letting the available NSFG data provide the most information. Prior studies comparing interviewer variance components for similar count variables could also be used in general applications of this technique to specify more informative prior 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 (example WinBUGS code used for the analyses is available in the Appendix).

Table 4.1 presents descriptive statistics for the interviewers in each of the groups defined by the eight interviewer-level characteristics. These descriptive statistics include the number of interviewers in each group (out of 87 total), and the mean, standard deviation (SD) and range for the number of cases (sample sizes) assigned to each interviewer.

**Table 4.1**
**Descriptive statistics for the NSFG interviewers in each group defined by the eight interviewer-level characteristics**

| | Number of Interviewers | Total Sample Size | Mean Sample Size | SD of Sample Sizes | Range of Sample Sizes |
|---|---|---|---|---|---|
| **Age (Years)** | | | | | |
| < 54 | 46 | 5,888 | 128.00 | 113.29 | (18, 554) |
| 55+ | 41 | 5,721 | 139.54 | 132.67 | (12, 532) |
| **Experience** | | | | | |
| < 5 Years | 49 | 6,062 | 123.71 | 126.65 | (12, 554) |
| 5+ Years | 38 | 5,547 | 145.97 | 116.71 | (18, 507) |
| **No. of Children** | | | | | |
| < 2 | 58 | 7,756 | 133.72 | 113.28 | (18, 532) |
| 2+ | 29 | 3,853 | 132.86 | 140.53 | (12, 554) |
| **Ever Married** | | | | | |
| Yes | 70 | 9,923 | 141.76 | 129.00 | (17, 554) |
| No | 17 | 1,686 | 99.18 | 83.49 | (12, 377) |
| **Other Job** | | | | | |
| No | 47 | 5,406 | 115.02 | 95.49 | (12, 532) |
| Yes | 40 | 6,203 | 155.08 | 145.92 | (17, 554) |
| **College Degree** | | | | | |
| No | 37 | 4,528 | 122.38 | 87.97 | (18, 409) |
| Yes | 50 | 7,081 | 141.62 | 142.71 | (12, 554) |
| **NSFG Before** | | | | | |
| No | 15 | 1,155 | 77.00 | 39.17 | (20, 166) |
| Yes | 72 | 10,454 | 145.19 | 130.29 | (12, 554) |
| **Ethnicity** | | | | | |
| Other | 16 | 1,781 | 111.31 | 75.53 | (20, 297) |
| White | 71 | 9,828 | 138.42 | 130.35 | (12, 554) |

The descriptive statistics in Table 4.1 indicate substantial variance in the sizes of the samples assigned to the interviewers. 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. We also note that for three of the observable interviewer features (Ever Married, NSFG Before, and Ethnicity), one of the two groups has less than 20 interviewers, which is not ideal for reliable estimation of variance components (Hox 1998). In light of the simulation results for smaller sample sizes (Section 3), we consider the impacts of these small sample sizes in our analyses.

Simple examinations of the distributions of the means of observed parity measures for the interviewers in each group are presented in Figure 4.1 below, to obtain an initial sense of the magnitude of interviewer variance in each of the groups. Figure 4.1 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 11,609 parity measures in the complete data set.

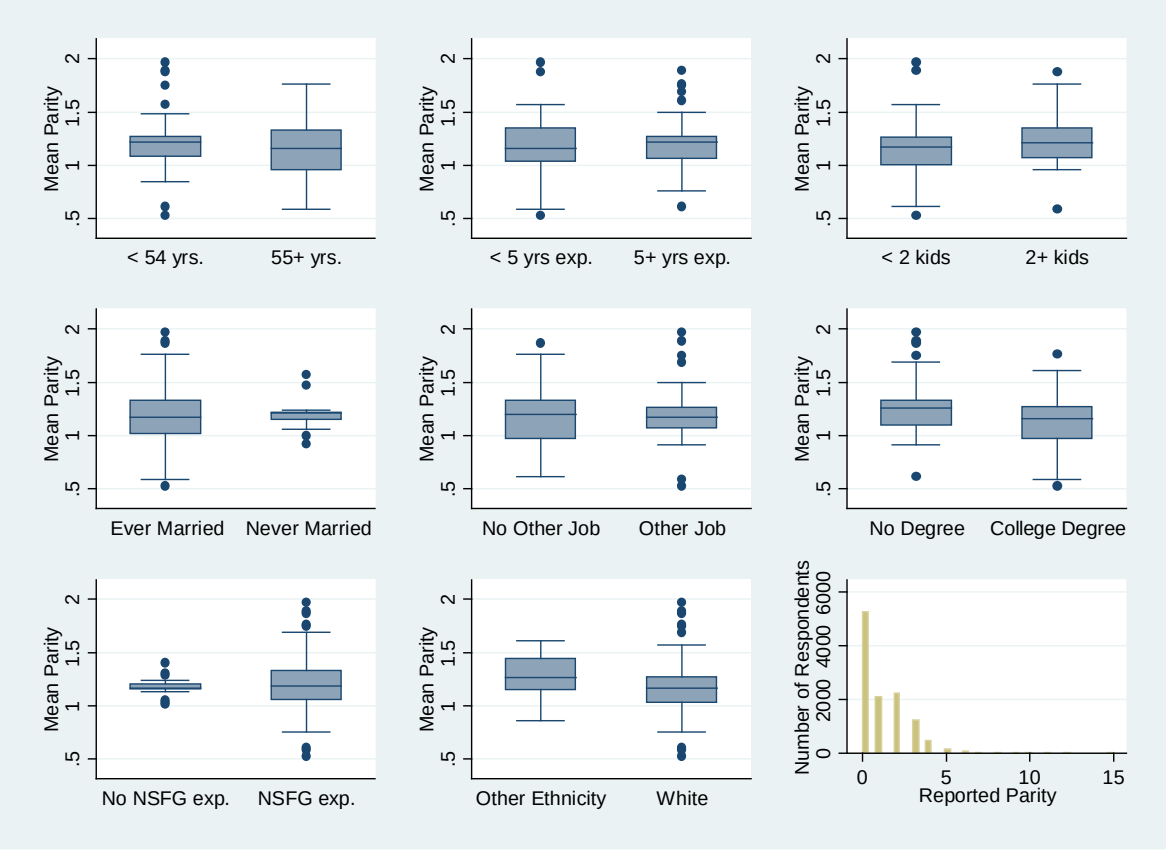


**Figure 4.1 Distributions of observed means on parity for interviewers in each group, with interviewer means weighted by assigned sample size, along with the overall distribution of the reported parity measures.**

The distributions of the means of measured parity values for the interviewers in Figure 4.1 provide an initial sense of groups that tend to differ in terms of the interviewer variance components. The group of interviewers that has never been married appears to have reduced variance, as does the group that has no

prior experience working on the NSFG. 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 do not produce different marginal means for the estimate of interest). Finally, the distribution of observed parity values for all 11,609 respondents has the expected appearance for a variable measuring a count of relatively rare events (live births), with mean 1.19 and variance 1.99.

We next consider the distributions of the proportions of females indicating current sexual activity among the interviewers in each group (Figure 4.2).

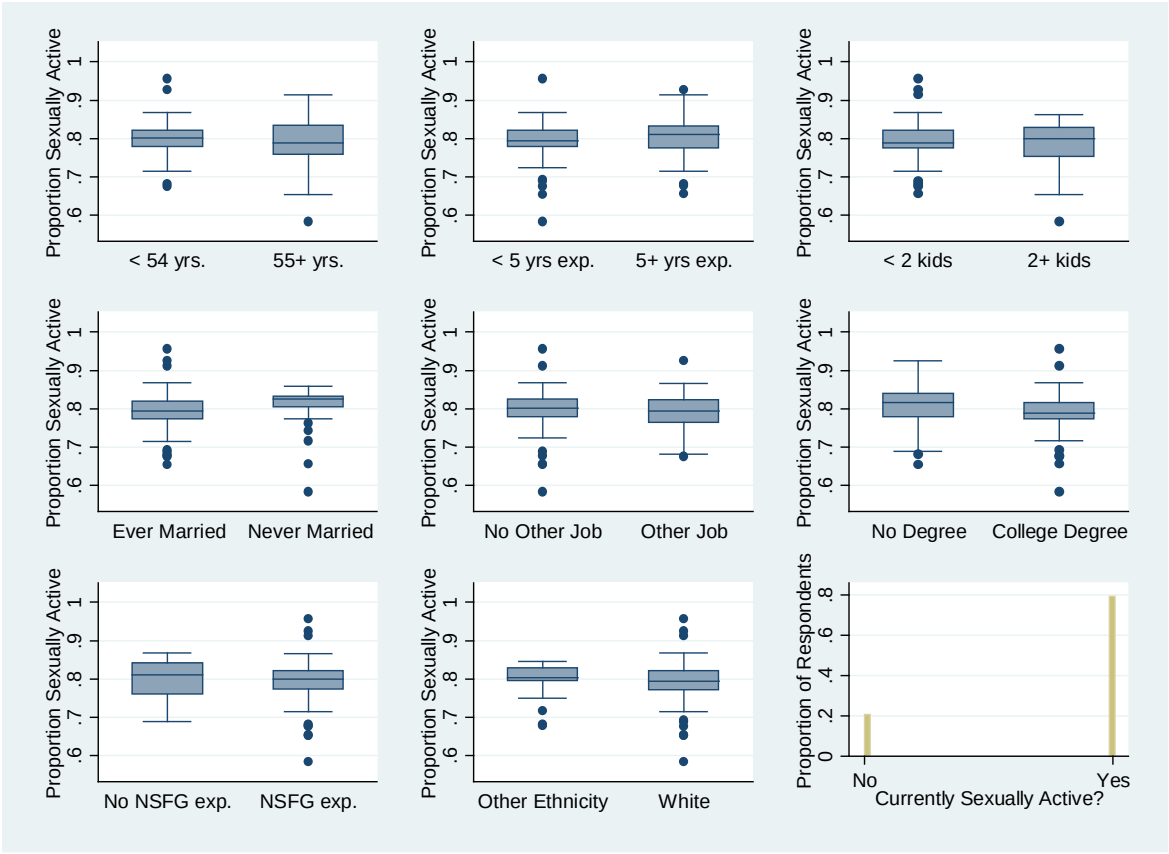


**Figure 4.2 Distributions of observed proportions of female respondents indicating current sexual activity for interviewers in each group, with interviewer means weighted by assigned sample size, along with the overall distribution of the sexual activity indicator.**

We see less evidence of differences in interviewer variance between the groups in general for this proportion, relative to average parity. Approximately 80% of the female respondents indicated that they were currently in a sexually active relationship.

Table 4.2 presents estimates of the parameters in each of the negative binomial models for the measured parity variable based on the two alternative analytic approaches. This table also presents results of the likelihood ratio tests comparing the two interviewer variance components (for each pair of groups) when following the frequentist approach, and 95% credible sets for the difference in the two variance components when following the Bayesian approach.

**Table 4.2**
**Parameter estimates in the negative binomial regression models for parity and comparisons of the interviewer variance components following the alternative frequentist and Bayesian analytic approaches.**

| | Frequentist Approach (SAS PROC GLIMMIX) | | | | Bayesian Approach (WinBUGS) | | | |
|---|---|---|---|---|---|---|---|---|
| **Interviewer Group Variable** | $\hat{\beta}_0$ **(SE)/** $\hat{\beta}_1$ **(SE)** | $\hat{\alpha}$ **(SE)** | $\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)** | **Likelihood Ratio Test:** $\tau_1^2 = \tau_2^2$ | $\hat{\beta}_0$ **(SD)/** $\hat{\beta}_1$ **(SD)** | $\hat{\alpha}$ **(SD)** | $\hat{\tau}_1^2$ **(SD)/** $\hat{\tau}_2^2$ **(SD)** | **95% CS:** $\tau_1^2 - \tau_2^2$ |
| **Age** | | | | | | | | |
| (1 = <54 years, | 0.185(0.031)/ | 0.538 | 0.026(0.009)/ | $\chi_1^2$ =0.03, | 0.183(0.033)/ | 0.685 | 0.025(0.010)/ | (-0.026, |
| 2 = 55+ years) | -0.007(0.043) | (0.018) | 0.024(0.008) | p= 0.873 | -0.003(0.046) | (0.024) | 0.024(0.009) | 0.028) |
| **Experience** | | | | | | | | |
| (1 = <5 years, | 0.201(0.033)/ | 0.537 | 0.024(0.008)/ | $\chi_1^2$ =0.04, | 0.197(0.034)/ | 0.694 | 0.024(0.009)/ | (-0.032, |
| 2 = 5+ years) | -0.036(0.044) | (0.018) | 0.027(0.010) | p= 0.835 | -0.031(0.045) | (0.027) | 0.027(0.011) | 0.024) |
| **Number of Kids** | | | | | | | | |
| (1 = <2, | 0.254(0.036)/ | 0.537 | 0.023(0.007)/ | $\chi_1^2$ =0.01, | 0.253(0.038)/ | 0.692 | 0.023(0.007)/ | (-0.032, |
| 2 = 2+) | -0.109(0.044) | (0.018) | 0.022(0.009) | p= 0.926 | -0.109(0.045) | (0.025) | 0.023(0.012) | 0.024) |
| **Ever Married** | | | | | | | | |
| (1 = Yes, | 0.184(0.029)/ | 0.537 | 0.030(0.008)/ | $\chi_1^2$ =5.41, | 0.181(0.037)/ | 0.694 | 0.030(0.008)/ | (0.002, |
| 2 = No) | -0.001(0.039) | (0.018) | 0.000(N/A)* | p= 0.020 | 0.004(0.045) | (0.025) | 0.003 (0.007) | 0.048) |
| **Other Job** | | | | | | | | |
| (1 = Yes, | 0.186(0.031)/ | 0.538 | 0.022(0.009)/ | $\chi_1^2$ =0.15, | 0.188(0.032)/ | 0.688 | 0.020(0.010)/ | (-0.036, |
| 2 = No) | -0.009(0.043) | (0.018) | 0.027(0.008) | p= 0.699 | -0.010(0.044) | (0.025) | 0.028(0.010) | 0.021) |
| **College Degree** | | | | | | | | |
| (1 = Yes, | 0.242(0.031)/ | 0.538 | 0.023(0.008)/ | $\chi_1^2$ < 0.01, | 0.240(0.032)/ | 0.693 | 0.024(0.009)/ | (-0.025, |
| 2 = No) | -0.108(0.042) | (0.018) | 0.022(0.008) | p= 0.963 | -0.106(0.044) | (0.024) | 0.021(0.010) | 0.030) |
| **NSFG Before** | | | | | | | | |
| (1 = Yes, | 0.174(0.035)/ | 0.537 | 0.031(0.008)/ | $\chi_1^2$ = 8.26, | 0.169(0.036)/ | 0.692 | 0.030(0.008)/ | (0.006, |
| 2 = No) | 0.010(0.043) | (0.018) | 0.000(N/A)* | p= 0.004 | 0.013(0.045) | (0.026) | 0.001(0.005) | 0.050) |
| **Ethnicity** | | | | | | | | |
| (1 = White, | 0.217(0.046)/ | 0.537 | 0.027(0.007)/ | $\chi_1^2$ =0.38, | 0.220(0.051)/ | 0.690 | 0.026(0.008)/ | (-0.045, |
| 2 = Other) | -0.044(0.052) | (0.018) | 0.018(0.011) | p= 0.536 | -0.050(0.058) | (0.025) | 0.020(0.017) | 0.027) |

* PROC GLIMMIX indicated that the estimated variance-covariance matrix of the random effects was not positive definite, and the estimate was set to zero because the RPL estimate of the variance component was negative. The same result occurred when using adaptive quadrature instead of RPL.

Notes: Estimates following Bayesian approach are medians of draws from posterior distributions. SE = Asymptotic SE. SD = SD of draws from posterior distribution. CS = Credible Set.

Consistent with our simulation study in Section 3, the results in Table 4.2 show that it is not uncommon for the frequentist approach to yield negative estimates of interviewer variance components (which causes SAS PROC GLIMMIX to set the estimates equal to zero, and not report estimated standard errors for the estimates), especially for groups with smaller samples of interviewers. In two cases, this results in a significant likelihood ratio test statistic, which would suggest that the two variance components are different. In contrast, the Bayesian approach produces very small estimates of the variance components, and a 95% credible set for the difference in the variance components. For example, in the cases of marital status and prior NSFG experience, we see estimates that are consistent with Figure 4.1, suggesting that there is significantly lower variance in the parity measures among the never-married group of interviewers and the inexperienced group of interviewers. The credible sets for the differences in these two cases agree with the frequentist tests, but the lower limits of these sets are very close to zero, suggesting that the differences, while significant, may not be very strong. We view this as an advantage of the Bayesian approach.

The Bayesian approach yields only slightly larger standard errors (or posterior standard deviations) for the parameter estimates in nearly all cases, reflecting the uncertainty in the parameter estimates that is accounted for by the prior distributions. The use of non-informative priors in this case, which would result in a posterior distribution that is dominated by the likelihood function, is the likely reason for the similarity in these measures of uncertainty, and more informative priors may increase the efficiency of the Bayesian estimates. Estimates of the individual parameters and corresponding inferences about them are generally quite similar when following the two approaches, as suggested by the literature in Section 2, and the estimated fixed effects suggest that the different groups of interviewers do not have a tendency to collect different measures on the parity variable. Interestingly, both approaches agree that interviewers with fewer children and/or a four-year college degree have a tendency to collect lower measures on the parity variable, but these differences could certainly be due to other covariates not accounted for in these analyses. Finally, we see slightly different estimates of the negative binomial scale parameter when following the two approaches. This is to be expected, as the Bayesian approach uses the medians of posterior distributions while the frequentist approach uses the modes of likelihood functions. In addition, the posterior distributions are not exactly equal to the likelihood functions when proper priors are utilized. The frequentist estimates of the scale parameter were much closer to the Bayesian estimates when using adaptive quadrature with five quadrature points to approximate the negative binomial likelihoods (results not shown); frequentist inferences for the other parameters did not change when using this alternative estimation method.

We repeated these analyses for the binary indicator of current sexual activity. Table 4.3 presents the estimated parameters in the multilevel logistic regression models following each of the two approaches. Consistent with Figure 4.2, these analyses reveal no evidence of differences between the various groups of interviewers in the variance components or the expected values of this outcome. Inferences were once again quite similar when following the two approaches, and the variances of the estimated variance components were once again slightly larger when following the Bayesian approach.

**Table 4.3**
**Parameter estimates in the logistic regression models for current sexual activity and comparisons of the interviewer variance components following the alternative frequentist and Bayesian analytic approaches.**

| | Frequentist Approach (SAS PROC GLIMMIX) | | | Bayesian Approach (WinBUGS) | | |
|---|---|---|---|---|---|---|
| **Interviewer Group Variable** | $\hat{\beta}_0$ **(SE)/** $\hat{\beta}_1$ **(SE)** | $\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)** | **Likelihood Ratio Test:** $\tau_1^2 = \tau_2^2$ | $\hat{\beta}_0$ **(SD)/** $\hat{\beta}_1$ **(SD)** | $\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)** | **95% CS:** $\tau_1^2 - \tau_2^2$ |
| **Age** (1 = <54 years, 2 = 55+ years) | 1.333 (0.066) / 0.032 (0.076) | 0.008 (0.013) / 0.045 (0.024) | $\chi_1^2 = 2.05$, $p = 0.153$ | 1.344 (0.055) / 0.024 (0.066) | 0.009 (0.013) / 0.046 (0.028) | (-0.107, 0.016) |
| **Experience** (1 = <5 years, 2 = 5+ years) | 1.378 (0.064) / -0.050 (0.073) | 0.004 (0.017) / 0.037 (0.020) | $\chi_1^2 = 1.52$, $p = 0.217$ | 1.384 (0.051) / -0.061 (0.064) | 0.005 (0.017) / 0.039 (0.023) | (-0.087, 0.024) |
| **Number of Kids** (1 = <2, 2 = 2+) | 1.362 (0.080) / -0.015 (0.088) | 0.022 (0.015) / 0.033 (0.024) | $\chi_1^2 = 0.16$, $p = 0.689$ | 1.363 (0.059) / -0.012 (0.070) | 0.024 (0.016) / 0.037 (0.030) | (-0.094, 0.037) |
| **Ever Married** (1 = Yes, 2 = No) | 1.387 (0.130) / -0.045 (0.134) | 0.020 (0.012) / 0.048 (0.041) | $\chi_1^2 = 0.58$, $p = 0.447$ | 1.398 (0.090) / -0.053 (0.097) | 0.021 (0.013) / 0.051 (0.055) | (-0.180, 0.035) |
| **Other Job** (1 = Yes, 2 = No) | 1.374 (0.043) / -0.046 (0.072) | 0.026 (0.016) / 0.024 (0.020) | $\chi_1^2 = 0.01$, $p = 0.927$ | 1.381 (0.045) / -0.051 (0.065) | 0.029 (0.019) / 0.022 (0.022) | (-0.055, 0.063) |
| **College Degree** (1 = Yes, 2 = No) | 1.388 (0.051) / -0.063 (0.071) | 0.016 (0.014) / 0.035 (0.022) | $\chi_1^2 = 0.60$, $p = 0.439$ | 1.394 (0.052) / -0.072 (0.064) | 0.014 (0.016) / 0.038 (0.024) | (-0.079, 0.033) |
| **NSFG Before** (1 = Yes, 2 = No) | 1.363 (0.103) / -0.012 (0.111) | 0.020 (0.012) / 0.069 (0.055) | $\chi_1^2 = 1.20$, $p = 0.273$ | 1.381 (0.113) / -0.024 (0.118) | 0.021 (0.013) / 0.083 (0.084) | (-0.301, 0.019) |
| **Ethnicity** (1 = White, 2 = Other) | 1.354 (0.077) / -0.004 (0.088) | 0.024 (0.014) / 0.032 (0.031) | $\chi_1^2 = 0.05$, $p = 0.816$ | 1.365 (0.080) / -0.013 (0.088) | 0.025 (0.015) / 0.032 (0.044) | (-0.131, 0.044) |

Notes: Estimates following Bayesian approach are medians of draws from posterior distributions. SE = Asymptotic SE. SD = SD of draws from posterior distribution. CS = Credible Set.

# 5 Concluding Remarks

This paper has considered frequentist and Bayesian methods for comparing the interviewer variance components for non-normally distributed survey items between two independent groups of survey interviewers. The methods are based on a flexible class of hierarchical generalized linear models (HGLMs) that allow the variance components for two mutually exclusive groups of interviewers to vary, and alternative inferential approaches based on those models. Results from a simulation study suggest that the two approaches have little empirical bias, comparable empirical MSE values and good coverage for moderate-to-large samples of interviewers and respondents. Analyses of real data from the U.S. National Survey of Family Growth (NSFG) suggest that inferences based on the two approaches tend to be quite similar. We find the similar performance of these two approaches to be good news for survey researchers, in that frequentists and Bayesians alike have tools available to them for analyzing this problem that will lead to similar conclusions.

There are some subtle distinctions between the two approaches that emerged in the analyses, mainly related to sample sizes and estimates of variance components that are extremely small or equal to zero.

These issues warrant further discussion, given their implications for survey practice. The Bayesian approach illustrated here is capable of accommodating uncertainty in the estimation of variance components when forming credible sets and does not rely on asymptotic theory, but we found that inferences about differences in variance components between a number of different subgroups of NSFG interviewers (each of moderate size) did not vary from those that would be made using frequentist approaches. Whether or not we would see the same results for even smaller groups of interviewers requires future investigation; the simulation study presented in Section 3 suggested that neither method performs well in a context where two groups of 20 interviewers collect data from 10 respondents each. An initial application of these two methods to data from the first quarter of data collection in this cycle of the NSFG (with about 20 interviewers in each of two groups interviewing about 20 respondents each on average) yielded findings similar to those reported here for larger samples, with some evidence of the Bayesian approach being more conservative (West 2011).

In general, the Bayesian approach provides a more natural form of inference for this problem, indicating a range of values for the difference in which approximately 95% of differences will fall. This may appeal to certain consumers of a given survey's products, as opposed to the simple $p$-value for a likelihood ratio test, which does not give users a sense of the range of possible differences. In the frequentist setting, the likelihood ratio test may be the only method of inference available if the pseudo maximum likelihood point estimate for one or more of the variance components is zero, with no corresponding standard error (preventing computation of Wald-type intervals). This situation was observed in both the simulations and the NSFG analyses, especially for groups with smaller samples of interviewers; given the reliance of likelihood ratio tests on asymptotic theory, the Bayesian approach may be better a better choice for smaller samples. The performance of the Bayesian approach is not ideal, however, for very small samples, as illustrated in the simulation study in Section 3.

We noted two significant differences between subgroups of interviewers in the NSFG data, and in each of these cases, the group with the smaller variance had an estimated variance component set to zero (with no standard error computed) when using the frequentist approach. The resulting inferences based on these estimates (where likelihood values were computed using the estimates of zero for the subgroups in question when performing the likelihood ratio tests) agreed with the Bayesian approach. We remind readers using frequentist methods that small samples of interviewers or extremely small amounts of variance among interviewers for particular variables may lead to negative maximum likelihood estimates of variance components, which can be problematic for the interpretation of interviewer variance for individual groups. Some software procedures capable of fitting multilevel models (e.g., the gllamm procedure in Stata, or the lmer() function in R) constrain variance components to be greater than zero during estimation to prevent this problem, which can increase estimation times. Other software procedures (like GLIMMIX in SAS) will simply fix these negative estimates to be zero, and fail to compute an estimated standard error. While these variance components technically cannot be equal to zero, we suggest interpreting these findings as evidence that there in negligible variance among the interviewers in a particular group. Bates (2009) argues against the use of standard errors for making inferences about variance components in the frequentist setting, especially when variance components are close to zero, instead suggesting that the profiled deviance function should be used to visualize the precision of the estimates. Both this approach and the Wald approach to computing confidence intervals will still be limited by smaller samples.

We do not see an empirical problem with using these zero estimates to perform the likelihood ratio tests demonstrated here for comparing groups of interviewers, given that Bayesian draws of the variance components in these groups would also be very small. However, in the case of estimating interviewer variance for single groups, examination of the sensitivity of Bayesian inferences to choices of different prior distributions for the variance components should be performed when variance components close to zero are expected, or the number of interviewers is relatively small (Browne and Draper 2006; Lambert et al. 2005). Furthermore, if survey researchers are interested in *predicting* random interviewer effects in the case where interviewer variance components are expected to be close to zero, both frequentist and Bayesian methods perform very poorly, and prediction is not recommended in this case (Singh et al. 1998, p. 390). See Savalei and Kolenikov (2008) for more discussion of the zero variance issue.

This study was certainly not without limitations. We acknowledge that the design of the NSFG, where interviewers are typically assigned to work in a single primary sampling area, did not allow for interpenetrated assignment of sampled cases to interviewers. As a result, disentangling interviewer effects from effects of the primary sampling areas is difficult. The methodologies illustrated in this paper can easily incorporate additional interviewer- or area-level covariates in an effort to "explain" variance among interviewers or areas due to observable covariates. The question of how to estimate interviewer variance in the presence of a strictly non-interpenetrated sample design needs more research in general, and we did not address this open question in this paper. As mentioned in Section 1, interpenetrated sample designs have been used in recent studies to disentangle interviewer and area effects. Future studies should examine the ability of the two approaches reviewed in this paper to detect differences in interviewer variance components when using cross-classified multilevel models that also include the effects of areas in an interpenetrated sample design.

On a similar note, we did not account for any of the complex sampling features of the NSFG (i.e., weighting or stratified cluster sampling) in the analyses. The theory that underlies the estimation of parameters in multilevel models in the presence of survey weights calls for weights for both the respondents and the higher-level clusters, which in this case would be interviewers (Rabe-Hesketh and Skrondal 2006; Pfefferman, Skinner, Holmes, Goldstein and Rasbash 1998). The analyses presented here effectively assume that we have a sample of interviewers from some larger population that was selected with equal probability, and that all respondents within each interviewer had equal weight. Methods outlined by Gabler and Lahiri (2009) might prove useful for addressing this limitation, and analysts could also include fixed effects of survey weights or stratification codes in the models proposed here. We leave these extensions for future research.

Finally, this paper also did not consider another rich aspect of the Bayesian approach, in that posterior draws of the 87 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.

# Acknowledgements

The authors are grateful for support from a contract with the National Center for Health Statistics that enabled the seventh cycle of the National Survey of Family Growth (contract 200-2000-07001).


# Appendix

## A.1 Example Code

We provide example code for fitting the types of models discussed in the paper using SAS PROC GLIMMIX below. In this code, PARITY and SEXMAIN are the count and binary variables, respectively, measured on NSFG respondents, FINAL_INT_ID is a final interviewer ID code, and INT_NVMARRIED is an indicator variable for whether or not an interviewer has never been married. The ASYCOV option will print asymptotic estimates of the variances and covariances of the estimated variance components.

```
/* marital status */

proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model parity = int_nvmarried / dist = negbin link = log solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;

proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model sexmain (event = "1") = int_nvmarried / dist = binary link = logit
solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;
```

We also provide example WinBUGS code for fitting the models using the Bayesian approaches discussed below. We call the WinBUGS code from the R software. In this code, LOWAGE.G is an interviewer-level indicator (with 87 values) for being in the younger interviewer age group, and HIGHAGE.G is an indicator for being in the older group. The full code, including code creating the variables used below, is available from the authors upon request.

```
# load necessary packages for using BUGS from R

library(arm)
library(R2WinBUGS)

############## Parity Analyses
```

```
# BUGS file for Age Group and Parity (age_nb.bug)

model {
   for (i in 1:n){
      parity[i] ~ dpois(lambda[i])
      lambda[i] <- rho[i]*mu[i]
      log(mu[i]) <- b0[intid[i]]
      rho[i]~dgamma(alpha,alpha)
   }

   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }

   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)
   alpha <- exp(logalpha)
   logalpha ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations for Parity/Age Group model in BUGS

n <- length(parity)
J <- 87
age.data <- list("n", "J", "parity", "intid", "highage.g", "lowage.g")
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2),
logalpha=rnorm(1))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0", "alpha")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_nb.bug", n.chains = 3,
n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

attach.bugs(age.1)

# for tables of results and inference

resultsmat <- cbind(numeric(6),numeric(6),numeric(6),numeric(6))

resultsmat[1,1] <- quantile(beta0,0.5)
resultsmat[1,2] <- sd(beta0)
resultsmat[1,3] <- quantile(beta0,0.025)
resultsmat[1,4] <- quantile(beta0,0.975)

resultsmat[2,1] <- quantile(beta1,0.5)
resultsmat[2,2] <- sd(beta1)
resultsmat[2,3] <- quantile(beta1,0.025)
resultsmat[2,4] <- quantile(beta1,0.975)

resultsmat[3,1] <- quantile(sigma.b0[,1]^2,0.5)
resultsmat[3,2] <- sd(sigma.b0[,1]^2)
```

```
resultsmat[3,3] <- quantile(sigma.b0[,1]^2,0.025)
resultsmat[3,4] <- quantile(sigma.b0[,1]^2,0.975)

resultsmat[4,1] <- quantile(sigma.b0[,2]^2,0.5)
resultsmat[4,2] <- sd(sigma.b0[,2]^2)
resultsmat[4,3] <- quantile(sigma.b0[,2]^2,0.025)
resultsmat[4,4] <- quantile(sigma.b0[,2]^2,0.975)

resultsmat[5,1] <- quantile(1/alpha,0.5)
resultsmat[5,2] <- sd(1/alpha)
resultsmat[5,3] <- quantile(1/alpha,0.025)
resultsmat[5,4] <- quantile(1/alpha,0.975)

vardiff <- sigma.b0[,1]^2 - sigma.b0[,2]^2
resultsmat[6,1] <- quantile(vardiff,0.5)
resultsmat[6,2] <- sd(vardiff)
resultsmat[6,3] <- quantile(vardiff,0.025)
resultsmat[6,4] <- quantile(vardiff,0.975)

resultsmat

############## Current Sexual Activity Analyses

# BUGS file for Age Group and Sexual Activity (age_bin.bug)

model {
   for (i in 1:n){
      sexmain[i] ~ dbern(p[i])
      logit(p[i]) <- b0[intid[i]]
   }

   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }
   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations for Parity/Age Group model in BUGS

n <- length(sexmain)
J <- 87
age.data <- list("n", "J", "sexmain", "intid", "highage.g", "lowage.g")
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_bin.bug", n.chains =
3, n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

attach.bugs(age.1)
```

# References

Bates, D. (2009). Assessing the precision of estimates of variance components. *Presentation to the Max Planck Institute for Ornithology*, Seewiesen, July 21, 2009. Presentation can be downloaded from http://lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4PrecisionD.pdf.

Biemer, P.P. and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. Chapter 27 of *Survey Measurement and Process Quality*, Editors Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin. Wiley-Interscience, 603-632.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Browne, W.J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473-514.

BUGS, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.html.

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

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one-way random model. *Technometrics*, 29(3), 323-337.

Collins, M. and Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

Durham, C.A., Pardoe, I. and Vega, E. (2004). A methodology for evaluating how product characteristics impact choice in retail settings with many zero observations: An application to restaurant wine purchase. *Journal of Agricultural and Resource Economics*, 29(1), 112-131.

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

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

Farrell, P.J. (2000). Bayesian inference for small area proportions. Sankhya: The *Indian Journal of Statistics*, Series B (1960-2002), 62(3), 402-416.

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

Gabler, S. and Lahiri, P. (2009). On the definition and interpretation of interviewer variability for a complex sampling design. *Survey Methodology*, 35(1), 85-99.

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.

Goldstein, H. (1995). *Multilevel Statistical Models, Second Edition*. Kendall's Library of Statistics 3, Edward Arnold.

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

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge University Press.

Hox, J. (1998). *Multilevel Modeling: When and Why*. In I. Balderjahn, R. Mathar and M. Schader (Eds.). Classification, data anaylsis, and data highways. New York: Springer-Verlag, 147-154.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R. and Jones, D.R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15), 2401-2428.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M. and Van Hoewyk, J. (2010). The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey. National Center for Health Statistics, *Vital and Health Statistics*, 2(150), June 2010.

Lynn, P., Kaminska, O. and Goldstein, H. (2011). Panel attrition: how important is it to keep the same interviewer? *ISER Working Paper Series*, Paper 2011-02.

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

Mangione, T.W., Fowler, F.J. and Louis, T.A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, 8(3), 293-307.

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.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162(3), 437-446.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60(1), 23-40.

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.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society*, Series A, 169, 805-827.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society, Series A*, 164(2), 339-355.

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

Savalei, V. and Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13(2), 150-170.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). *Handbook of Survey Research, Second Edition*. In Interviewers and Interviewing. J.D. Wright and P.V. Marsden (eds). Bingley, U.K.: Emerald Group Publishing Limited.

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

Schober, M. and Conrad, F. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, 60(2), 377-396.

Ugarte, M.D., Goicoa, T. and Militino, A.F. (2009). Empirical bayes and fully bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics and Data Analysis*, 53, 2938-2949.

Van Tassell, C.P. and Van Vleck, L.D. (1996). Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co)variance component inference. *Journal of Animal Science*, 74, 2586-2597.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26, 37-52.

West, B.T. (2011). Bayesian analysis of between-group differences in variance components in hierarchical generalized linear models. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: *American Statistical Association*, 1828-1842.

West, B.T. and Galecki, A.T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician*, 65(4), 274-282.

West, B.T., Kreuter, F. and Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29(2), 277-297.

West, B.T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5), 1004-1026.

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.