Hierarchical Bayes Small Area Estimates of Adult Literacy Using Unmatched Sampling and Linking Models

Leyla Mohadjer[1], J.N.K. Rao[2], Benmei Liu[1], Tom Krenzke[1], and Wendy Van de Kerckhove[1]
Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, MD 20850[1]
School of Mathematics and Statistics, Carleton University, Ottawa, Canada, KIS 5B6[2]

Abstract

Funded by the National Center for Education Statistics, the National Assessment of Adult Literacy (NAAL) was designed to measure the English literacy skills of adults in the U.S. based on an assessment containing a series of literacy tasks completed by sampled adults. Sufficiently precise estimates have been produced for the nation and major subdomains of interest using the NAAL data. However, policymakers and researchers/business leaders often need literacy information for states and counties but these areas do not have large enough samples to produce reliable estimates. Therefore, small area estimation techniques are used to produce estimates of literacy levels for all states and counties in the nation. This paper describes the Hierarchical Bayesian estimation techniques used to derive a single area-level linking model to produce both county and state estimates, and credible intervals.

Keywords: Indirect state and county estimates, credible intervals, variance smoothing, predictor variable selection

1. Introduction

This paper describes the statistical methodology used to produce small area county and state estimates of the percent of adults at the lowest literacy level based on survey data from the 2003 National Assessment of Adult Literacy (NAAL), sponsored by the National Center for Education Statistics (NCES). The paper describes the steps taken in the development and evaluation of the small area model. The small area estimates are currently under review by NCES and are not discussed in this paper.

The 2003 NAAL was designed to measure the nation's English literacy skills. The NAAL conducted interviews with a sample of adults residing in private households in the United States. The sample represents the household population of U.S. adults who were age 16 and older, from the 50 states and the District of Columbia. Over 18,500 adults participated in the household component of NAAL. The survey was conducted from May 2003 through February 2004 and was made up of a national sample of adults supplemented by state samples in six states (Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma) that participated in the State Assessment of Adult Literacy (SAAL). NAAL was also designed to provide high-precision estimates for Blacks and Hispanics. To accomplish this, oversampling was carried out for these two subgroups in the national sample. In addition to the household component, approximately 1,200 inmates of federal and state prisons were assessed. The inmate sample does not contribute to the NAAL small area estimates.

Each individual who participated in the NAAL provided demographic and other background information, and was asked to complete a booklet containing a series of literacy tasks. The tasks measured each individual's ability to use printed and written information to function in society on the basis of three literacy scales: Prose, Document, and Quantitative literacy. A set of booklets containing different sets of tasks was used so that the sampled individuals did not all perform the same tasks. Item Response Theory (IRT) methods were used to create the three scales. Four categories were established to describe the literacy levels for each scale: Below Basic, Basic, Intermediate, and Proficient. The NAAL reports (see http://nces.ed.gov/NAAL) provide results for the literacy levels of adults for each of the three scales separately.

The NAAL sample size is large enough to provide estimates of literacy levels for the nation and for major subdomains of interest that are sufficiently precise. In addition, states that participated in the SAAL are guaranteed reliable estimates of literacy levels for the three scales for their states and their major subdomains. However, other states and jurisdictions within states such as counties, do not have large enough sample sizes to produce estimates of adequate precision (some larger states may have sufficient sample sizes, but the NAAL design does not support state-level estimation). Indeed, some states and most counties have no sample in the NAAL. Nevertheless, policymakers, business leaders, and educators/researchers often need literacy information for states and counties. In response, NCES has used statistical modeling approaches to produce model-dependent estimates of the percentages of adults in the lowest literacy level (those who could not be tested because of language barriers and those at the Below Basic level) on the prose scale for all states and counties in the nation. These estimates are called “indirect” estimates to distinguish them from standard survey or “direct” estimates that do not depend on the validity of a statistical model. The indirect estimates are produced using small area estimation techniques that rely on survey data from other areas and auxiliary data for the area obtained from other sources (such as the decennial census) to “borrow strength” in creating the estimates.

As mentioned above, IRT modeling was used for creating literacy scores for the NAAL participants. Two aspects of the NAAL IRT modeling needed to be taken into account when developing the small area modeling approach. First, direct estimates produced for subgroups are based on IRT models that are different from the model used for the aggregate group. This implies that state direct estimates cannot be produced by combining the estimates for all the counties in the state. However, further examinations showed negligible differences between the direct state and aggregated county estimates given the precision levels of the associated direct and indirect estimates. Therefore, indirect state estimates were computed as the aggregated indirect county estimates.

Second, the variances of the IRT-based estimates of literacy proficiency not only reflect the survey sampling error, but also measure the variances coming from the IRT estimation approach. Some evaluations were carried out to examine these components of the direct variances prior to developing the small area modeling approach for the NAAL.

Section 2 describes the Hierarchical Bayesian (HB) estimation technique used to create a single area-level model for producing the state and county-level estimates. The section includes a description of the numerous state- and county-level auxiliary variables considered as predictor variables for use in the small area model. It also describes the methodology used to select the set of variables chosen for the final model and lists the six predictor variables included in the final model. The small area modeling approach used for the NAAL assumes that the relative variances (or relvariances) of the direct estimates are known. In practice, only highly imprecise estimates of the relvariances are known. Section 2 also includes a description of the methodology used to smooth the direct variances. Section 3 provides a description of the approaches used in fitting the final model, and Section 4 describes the approaches used to produce estimates for counties with sample data, for counties with no data, and for states. In addition, a description of how the credible intervals were computed for all the NAAL indirect estimates is included, followed by a description of methods used to conduct comparisons between pairs of counties and states. Section 5 provides a brief summary of various steps taken to evaluate the model and the indirect estimates. It also explains why benchmarking the county estimates to aggregate direct survey estimates was not employed. Finally, Section 6 provides a summary and conclusions.

2. Small Area Model Development

A single HB model has been used to produce both county and state indirect estimates of the percentages of adults at the lowest level of literacy. The model has two separate components: a sampling model and an unmatched linking model. These models are described in turn below. More details are provided in Chapter 10 of Small Area Estimation (Rao 2003).

2.1 Sampling Model

The sampling model is given by

$$p_{ij} = \theta_{ij} + \varepsilon_{ij} \tag{1}$$

where p_{ij} is the direct estimate and θ_{ij} is the true value of the proportion of adults at the lowest level of literacy in county *j* in state *i*. The model assumptions are that the error term ε_{ij} is normally distributed with a mean of 0 and a variance of ψ_{ij}, i.e., $\varepsilon_{ij} \sim N(0, \psi_{ij})$, and the HB model further assumes that the relvariance $\varphi_{ij}^2 = \psi_{ij} \big/ \theta_{ij}^2$ is known.

There are two aspects of this model that deserve comment. First, the normality assumption is somewhat problematic because the sample sizes in many counties are small and the values of θ_{ij} are also often fairly small. This assumption is required for the assumed HB model and follows the general practice in modeling small area estimates. Second, the assumption that φ_{ij}^2 is known does not hold, and moreover, the sample estimates for these relvariances are generally very imprecise. To address this issue, models have been developed to predict φ_{ij}^2, with the model predictions then being assumed to be the true values, again following the general practice adapted in these situations.

2.2 Linking Model

The purpose of the linking model is to relate the values of θ_{ij} to a set of auxiliary variables that are predictors of θ_{ij}. Since θ_{ij} is a proportion, a logit model is assumed:

$$\text{logit}(\theta_{ij}) = \sum_{k=1}^{K} \beta_k x_{ijk} + v_i + u_{ij} \qquad (2)$$

where $\text{logit}(\theta_{ij}) = \log[\theta_{ij} / (1-\theta_{ij})]$, x_{ijk} are a set of K-1 predictor variables and an intercept term (i.e., $x_{ij1} = 1$), the β_k are a set of regression coefficients, v_i is a state random effect ($v_i \overset{iid}{\sim} N(0, \sigma_v^2)$), and u_{ij} is a county random effect ($u_{ij} \overset{iid}{\sim} N(0, \sigma_u^2)$).

The following widely used prior distributions are assumed for the parameters on the right-hand side of the linking model:

Each of the β_k has a flat prior distribution;

$$\sigma_v^2 \sim ING(0.001,\ 0.001)\,,$$

where ING denotes the inverse gamma distribution; and

$$\sigma_u^2 \sim ING(0.001,\ 0.001)\,.$$

The combination of the sampling model and the linking model is termed an unmatched model because the two model components cannot be simply merged into a single model, as would be the case if the linking model had been a linear rather than a logit model. Model fitting consists of producing posterior distributions for all the model parameters:

$$\boldsymbol{\eta} = \left(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u}, \sigma_v^2, \sigma_u^2\right)$$

where boldface letters denote matrices or vectors of the associated multiple parameters.

2.3 Auxiliary Variables

A key aspect of small area estimation modeling for the NAAL was finding auxiliary variables that are measured consistently across all counties and states, and that are effective predictors of the percent of adults at the lowest literacy level. The importance of identifying literacy-related auxiliary data is magnified for NAAL, since about 10 percent of the counties in the United States contain NAAL sample. The remaining counties rely on auxiliary data from sources other than the NAAL survey. In addition, the auxiliary data help to improve the precision of estimates for counties that have NAAL sample.

Given the importance of finding good predictors, a considerable effort was devoted to identifying reliable data sources and variables that are potential predictors of literacy. In total, over 100 auxiliary variables across 20 major variable types (e.g., poverty, income, education, occupation, etc.) were obtained as potential predictors for the percent at the lowest literacy level. The primary source was county-level data from the 2000 Census of Population. The census data contains a wealth of variables, several of which, such as country of birth, education, age, and disabilities, have been known through past analyses to be related to adult literacy skills (see Kirsch et al. 1993 and Greenberg et al. 2001).

Once the set of auxiliary variables was accumulated, a two-phase variable selection process was implemented. In the first phase, the long lists of county and state-level auxiliary variables were reduced by retaining only the critical variables after a

1. Bivariate correlation analysis between the auxiliary variables and the percent at the lowest literacy level, and analytical model fitting;

2. Search of variables known to be correlated with literacy from past analyses or hypothesized to be correlated with literacy; and

3. Review of sample design variables with impact on small area modeling.

Once the lists of auxiliary variables were reduced, the second phase evaluated the variables using SAS Proc Mixed and WinBUGS[1]. The result of this process was the set of predictor variables retained for the small area modeling. The model included predictor variables relating to foreign-born status, education attainment, race/ethnicity, poverty status, census division indicator and state assessment indicator.

---

[1] For more information about the WinBUGS software, refer to http://www.mrc-bsu.cam.ac.uk/bugs/.

2.4 Smoothing Direct Relative Variances

The small area modeling approach used for estimating the percent of adults at the lowest literacy level assumes that the relvariances of the direct county estimates are known. In practice, only highly imprecise estimates of the relvariances are available. These estimates need to be "smoothed" and they are then assumed known.

Expanding on the methods described in Wolter (2007), a two-step approach was developed to produce model-dependent estimates of the relvariances. The two steps were based on the requirement that the predicted relvariances should not depend directly on the direct survey estimates or variance estimate.

In the first step, the proportion at the lowest literacy level was predicted from a weighted robust regression model relating the direct estimates of p_{ij} to auxiliary variables. Each county was assigned a weight of the square root of its sample size on the grounds that its sampling error—which was related to its sample size—was an important part of its residual error in the regression model. The predictor variables were selected using a stepwise selection method from the list of auxiliary variables chosen for the small area modeling, as described above. The final step 1 model had the form:

$$logit\left(p_{ij}\right) = \gamma_0 + \gamma_1 Z_{ij1} + \gamma_2 Z_{ij2} + \gamma_3 Z_{ij3} + \gamma_4 Z_{ij4} + \varepsilon_{ij} \qquad (3)$$

where p_{ij} is the proportion at the lowest literacy level; Z_{ij1} - Z_{ij4} are county-level auxiliary variables; and the error term ε_{ij} is normally distributed with a mean of 0 and a variance of σ^2.

In the second step, the predicted values of the proportions at the lowest level of literacy from the regression model in Equation (3) were used in a generalized variance function (GVF) model to smooth the relvariance estimates. This model draws on a sampling error model, where relvariances are functions of the predicted values from Equation 3 and the sample size in the county. To make the model linear in the parameters, a robust weighted least squares log-log model was used. The weight was equal to the square root of the sample size to reflect the precision of the estimates. The model has the form:

$$\log(\varphi_{ij}^2) = \eta_0 + \eta_1 \log(\tilde{p}_{ij}) + \eta_2 \log(1-\tilde{p}_{ij}) + \eta_3 \log(n_{ij}) + \varepsilon_{ij} \qquad (4)$$

where φ_{ij}^2 is the relvariance of the proportion of adults at the lowest level of literacy; $\tilde{p}_{ij}$ is the predicted proportion from Equation (3); n_{ij} is the sample size; and the error term ε_{ij} is normally distributed with mean 0 and variance σ^2. The predicted values from Equation (4) were then treated as known relvariances in the small area modeling stage.

3. 2003 NAAL Model Fitting

Model selection process started with a preliminary comparison of many different models with alternative sets of auxiliary variables. Then a selected set of models were chosen to go through an extensive evaluation process, as described in Section 5. This section describes the procedures employed to fit the final model with the six variables given in Section 2.3.

HB estimation techniques with noninformative prior distributions were used to model the relationship between the predictor variables and the direct county estimates (the dependent variable). Model fitting was carried out using a Markov Chain Monte Carlo (MCMC) method. The WinBUGS software (Lunn et al. 2000), which uses the Metropolis-Hastings (M-H) algorithm within the Gibbs sampler, was employed for this purpose. Three independent Markov Chains (hereinafter referred to as "runs")[2] were processed to facilitate the calculation of Monte Carlo standard errors (see Gelman and Rubin 1992; Rao 2003, p. 229).

The procedure started with three sets of initial values for $\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u}$, σ_v^2, and σ_u^2, corresponding to the three independent MCMC runs and then updated all the values of $\boldsymbol{\eta}$ repeatedly within each set. The initial values were drawn following these steps. First, maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}, \boldsymbol{v}$, and $\boldsymbol{u}$ were produced, along with their variances σ_v^2 and σ_u^2 by running a random effects regression model for predicting θ_{ij} using SAS Proc Mixed. The

[2] The Markov Chains are also referred to as "chains" or "sequences" in this context.

distributions of $\boldsymbol{\beta}, \boldsymbol{v}, \text{and}\, \boldsymbol{u}$ were assumed to be approximately normal. The MLE variances were varied by 10 percent and were used to derive three sets of normal distributions for the parameters σ_v^2 and σ_u^2. For each set, initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{v}^{(0)}$, and $\boldsymbol{u}^{(0)}$ were drawn from the normal distributions.

Given a set of initial values, each run was then processed separately. For the first iteration in a run, the value of one component of $\boldsymbol{\eta}^{(0)}$ was updated, then the next component was updated using the updated value of the first component and the initial values of the other components, then the third component was updated using the updated values of the first two components and the initial values of the remaining components, and so on. The run's second iteration started with the updated values of all components and repeated the process. The process was repeated 10,000 times, until convergence was assumed to have been reached. The iterations up to this point (called the burn-in period) were discarded.

After that point, 90,000 further iterations were produced. Since the results from neighboring iterations after burn-in are correlated, they were "thinned" by taking a systematic sample of one in 10 of them. Thus, over the three runs, 27,000 iterations remained. These 27,000 final iterations (referred to as MCMC samples) then simulated the posterior distributions of all the parameters in $\boldsymbol{\eta}$. The means of the parameter estimates across the 27,000 MCMC samples are the HB estimates of the parameters.

Note that, given the value of $\hat{\theta}_{ij}$ at a particular MCMC sample, the sampling variance ψ_{ij} is derived from the assumed known relvariance, as $\hat{\psi}_{ij} = \hat{\theta}_{ij}^2 \tilde{\varphi}_{ij}^2$. Hence, it also has a posterior distribution.

The WinBUGS software provides the potential scale reduction factor estimate $\hat{R}$ as a convergence diagnostic for each of the parameters in $\boldsymbol{\eta}$. This statistic is based on an analysis of variance decomposition of the total variance in the values produced by three runs of length 90,000 each after burn-in. If convergence is attained, in expectation, the value of $\hat{R}$ should be close to 1 (Rao 2003, pp. 229-230). A value of $\hat{R}$ much larger than 1 suggests that a larger number of iterations is required for burn-in. The values of $\hat{R}$ for the parameters $\boldsymbol{\beta}$, σ_v^2, and σ_u^2 are all near 1 (ranging from 1.001 to 1.005). The values of $\hat{R}$ for $\boldsymbol{v}, \boldsymbol{u}, \text{and}\, \boldsymbol{\theta}$ are also all near 1, ranging from 1.001 to 1.002. The Brooks-Gelman-Rubin plots (Brooks and Gelman 1998) were reviewed as a graphical display of $\hat{R}$ and were also useful in determining the number of iterations to burn-in.

Throughout the initial testing of models, several other plots generated by WinBUGS were also reviewed. A visual inspection of autocorrelation plots was conducted to determine the thinning amount and to check for independent iterations. Trace plots were also reviewed to check for independence and convergence. In addition, a density plot was used to help determine the number of iterations.

4. Small Area Estimates for Counties and States

In general, the value of $\theta_{ij}^{(b)}$ for MCMC sample *b* is obtained from

$$\text{logit}\left(\theta_{\text{ij}}^{(\text{b})}\right) = \boldsymbol{x}_{\boldsymbol{ij}}'\boldsymbol{\beta}^{(b)} + v_i^{(b)} + u_{ij}^{(b)} \tag{5}$$

Then for sampled counties, the posterior mean $\hat{\theta}_{ij}^{HB}$, which is also called the indirect estimate of county-level posterior proportion for sampled county j within state i, is produced as:

$$\hat{\theta}_{ij}^{HB} = \frac{\sum_{b=1}^{27,000} \theta_{ij}^{(b)}}{27,000} \tag{6}$$

For sampled counties, estimates of all the components on the right hand side of this equation are available. However, for all of the nonsampled counties, the values of $u_{ij}^{(b)}$ were not available, and for non-sampled counties in states without a sampled county, values of $v_i^{(b)}$ were not available either.

For nonsampled counties in states with one or more sampled counties, the estimated state effect was available from WinBUGS. For such counties, the estimate of $\theta_{ij}^{(b)}$ was computed from

$$\text{logit}\,(\theta_{ij}^{(b)}) = \boldsymbol{x}_{\boldsymbol{ij}}'\boldsymbol{\beta}^{(b)} + v_i^{(b)} + u_{ij(RD)}^{(b)} \tag{7}$$

where $u_{ij(RD)}^{(b)}$ is a random draw from $N(0,\sigma_u^{2(b)})$. For nonsampled counties in states with no sampled county, the estimate of $\theta_{ij}^{(b)}$ was computed from

$$\text{logit}\,(\theta_{ij}^{(b)}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}^{(b)} + v_{i(RD)}^{(b)} + u_{ij(RD)}^{(b)} \qquad (8)$$

where $v_{i(RD)}^{(b)}$ is a random draw from $N(0,\sigma_v^{2(b)})$ and $u_{ij(RD)}^{(b)}$ is a random draw from $N(0,\sigma_u^{2(b)})$. In both cases, once the set of 27,000 values of $\theta_{ij}^{(b)}$ was obtained, the posterior mean for nonsampled counties was computed using Equation (6).

The indirect estimates for states were computed as weighted aggregates of indirect county estimates, where the weights represent the proportion of the state's household population of adults aged 16 and over in each county.

4.1 Measures of Precision

It is important to take the prediction error in model-dependent indirect estimates into account in their interpretation. Often this error is substantial. The NAAL indirect estimates are no exception. The primary measure of precision reported for each NAAL state or county indirect estimate is its credible interval.

A credible interval is a posterior probability interval, used in Bayesian statistics for purposes similar to those of a confidence interval in frequentist statistics. A 95 percent credible interval is any interval whose probability under the posterior distribution is 0.95. The 95 percent credible intervals for both the county estimates $\hat{\theta}_{ij}^{HB}$ and the state estimates $\hat{\theta}_i^{HB}$ were computed by calculating the 2.5 percent (lower bound) and 97.5 percent (upper bound) quantiles of $\hat{\theta}_{ij}^{(b)}$ and $\hat{\theta}_i^{(b)}$, respectively, from the 27,000 MCMC samples that simulated the posterior distributions. Since these posterior distributions are skewed, the credible intervals are nonsymmetric around the estimate.

4.2 Comparisons Between Pairs of States and Counties

In principle, the MCMC procedures can be extended to provide credible intervals for the differences between any pair of counties or states. For each MCMC sample, the quantity $\left(\hat{\theta}_{ij}^{HB} - \hat{\theta}_{i'j'}^{HB}\right)$ is computed and the credible interval for the difference $\left(\hat{\theta}_{ij}^{(b)} - \theta_{i'j'}^{(b)}\right)$ is then derived from the resultant posterior distribution. In practice, in view of the enormous number of possible pairwise comparisons between counties across the nation (about 5 million), this procedure has been applied only for differences between any pair of states and between any pair of counties that are within the same state.

5. Model Evaluation

Alternative models were fit to the data to determine if the model results were sensitive either to the prior distributions used for modeling or to the set of auxiliary variables in the model. Once the final model was selected, three measures of model fit were computed to assess how well the model fit the data.

A sizable number of models, with alternative sets of predictor variables, were compared in the model selection process. All of the models contained a core set of five variables that had shown to be important (foreign-born status, education attainment, race/ethnicity, census division indicator and state assessment indicator). Some additional variables and versions thereof (continuous, square root of the percentage, and a dichotomous recode), were introduced into the models either because past research had found them to be correlated with literacy or it was thought that they might improve the predictions for nonsampled counties.

The Deviance Information Criterion (DIC) (Spiegelhalter et. al. 2002) was used to compare the fit of the alternative models. The DIC is a measure of goodness of fit that takes account of the number of parameters in the model (like an adjusted R^2). A smaller value of DIC indicates a better fit. In general, a rough guideline was used to rule out a model with a DIC that exceeds the DIC for another model by at least ten (BUGS 2004). This rule is analogous to the one for the Akaike Information Criterion (AIC) used for logistic regression models (Burnham and Anderson, 2004).

In addition to the DIC measure, alternative models were evaluated with respect to improving estimates for the nonsampled counties.

As a last step in the model selection process, the county weight (the inverse of the county's selection probability) was added to the final model as a predictor variable. The purpose of this addition was to check for possible improvements in the model fit by reflecting the counties' selection probabilities (in general, larger counties had higher chances of selection). However, the correlations between the indirect HB county estimates obtained from the models with and without the weight variable exceeded .999 for sampled counties and .998 for nonsampled counties. It was therefore decided not to include the county weight as a predictor variable in the final model.

Once the final model was selected, the following three variants of the prior distributions were examined:

- Changing the noninformative flat prior distributions for the regression coefficients $\boldsymbol{\beta}$ to informative normal priors with mean 0 and very large variances;

- Changing the inverse gamma prior distributions for the variances of the county and state random effects from ING (0.001, 0.001) to ING (0.0001, 0.001) and ING (0.0001, 0.0001) (here, "ING" denotes the Inverse Gamma Distribution); and

- Changing the inverse gamma prior distributions ING (0.001, 0.001) for the variances of the county and state random effects to noninformative flat priors.

The correlations between the set of indirect estimates from the final model and each of the sets of indirect estimates based on the above alternative scenarios of prior distributions ranged from 0.993 to 1.000, indicating that the final estimates are not sensitive to the choice of the prior distributions.

In addition, three measures were computed to assess the goodness of fit (Rao 2003, Chapter 10):

- Global measure that compares two discrepancy measures, one based on the difference between the indirect and direct county estimates, and the other based on the difference between the indirect estimates and estimates simulated from the posterior normal distributions for the HB county estimates;

- County-level measure computed as the proportion of the 27,000 MCMC samples that had a smaller simulated value (as opposed to direct estimates); and

- County-level measure that is computed as the difference between the mean of the simulated values and the direct estimate, divided by the standard deviation of the simulated values, where the mean and standard deviation of the simulated values are computed across the 27,000 MCMC samples.

The three diagnostic measures showed that the model fit the county-level data generally very well.

Another useful method for evaluating indirect estimates is to compare them with the corresponding direct estimates at some aggregate geographical level for which the direct estimates are reasonably reliable. By forming aggregates of the areas—termed henceforth "domains"—in a variety of ways (for instance, by region, by poverty level and by population size), the comparisons provide tests of the indirect estimates along a number of dimensions. The indirect county-level estimates were aggregated to a number of domains using county-level characteristics following the same approach used to create state estimates. In general, the estimates are close and the aggregated indirect estimates always fall within the 95 percent confidence intervals for the direct estimates.

An issue related to these comparisons is whether to benchmark the indirect estimates to conform to direct estimates for certain large domains. Benchmarking is often attractive because it provides indirect estimates that are consistent with published direct estimates. However, this does not apply to the NAAL situation because the published estimates exclude the language barrier cases, whereas for the indirect estimates they are included. Furthermore, as discussed above, the NAAL IRT approach used for obtaining direct estimates for domains produces estimates that do not exactly conform to the direct estimates that would be obtained by aggregating estimates based on county level IRT modeling. And lastly, the differences between the aggregated indirect estimates and the direct estimates are small and within the bounds of sampling error. For the above reasons, a decision was made not to use any benchmarking for the NAAL indirect estimates.

6. Summary and Conclusions

This paper described the statistical methodology used to produce the model-dependent—indirect—county and state estimates of the percentages of adults at the lowest literacy level for individual states and counties for the 2003 NAAL. A Hierarchical Bayes (HB) model was adopted using a Markov Chain Monte

Carlo (MCMC) method. The model was implemented using the WinBUGS software. The key component of the approach was to develop a model to predict county percentages of adults at the lowest literacy level based on the survey data and a set of auxiliary variables that were available and measured consistently for all counties. Within the model selection process, systematic approaches were developed to identify, gather, analyze and select final predictor variables. Also, several approaches were considered and evaluated while arriving at the procedure summarized in this paper for smoothing the direct relvariances.

Various techniques were used to evaluate the fit of the HB model to the observed data. None of them indicated appreciable problems with the final model used to produce the county and state indirect estimates. First, alternative models were constructed using different prior distributions and different sets of auxiliary variables. This analysis supported the choice of the final model and indicated that the indirect estimates were not sensitive to the variants of the model that were investigated. The final model also proved satisfactory with regard to several diagnostic tests of fit. Lastly, comparisons of direct estimates for a variety of domains defined along different dimensions with aggregations of the indirect county estimates for those domains showed a close correspondence in each case. These evaluation checks and many others not mentioned here, all support the model used in creating the NAAL county- and state-level estimates.

Acknowledgements

Graham Kalton led the technical aspects of this task. The authors are grateful to him for playing a major role in the development of the model, evaluation, and the production of the final estimates. This work was funded by the National Center for Education Statistics and conducted under the supervision of Sheida White, the NAAL project Officer, and Andrew Kolstad, the NAAL senior technical advisor.

References

Brooks, S.P.. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, pp.434-455.

BUGS (2004). The BUGS project –DIC. Supported by MRC Biostatistics Unit, Cambridge, UK. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/ dicpage.shtml.

Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods Research*.

Chen, M., Shao, Qi., and Ibraham, J.G. (2000). Monto Carlo Methods in Bayesian Computation. Springer-Verlag New York, Inc.

Chip. S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, pp.327-335.

Datta, G.S., Lahiri, P., Maiti, T., and Lu, K.L. (1999). Hierarchical Bayes Estimation of the unemployment rates for the stats of the U.S., *Journal of the American Statistical Association*, Vol.94, No.448, pp.1074-1082.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequence, *Statistical Science*, 7, pp.457-472.

Greenberg, E., Macías, R.F., Rhodes, D., and Chan, T. (2001). *English Literacy and Language Minorities in the United States*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kirsch, I. S., Jungeblut, A., Jenkins, L., and Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Princeton, NJ: Educational Testing Service.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility, Statistics and Computing, Vol.10, No.4, pp.325-337.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling procedures in the National Assessment for Educational Progress. *Journal of Educational Statistics*, *17*, 131-154.

Rao, J.N.K (2003). Small Area Estimation, New York: John Wiley and Sons.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, 64: 583-640.

Wolter, K.M. (2007). Introduction to Variance Estimation, New York: Springer.

You, Y, and Rao, J.N.K (2002a). Small area estimation using unmatched sampling and linking models, *The Canadian Journal of Statistics*, Vol.30, No.1, pp.3-15.