US Department of Education
NCES 2009-482

National Assessment of Adult Literacy

Indirect County and State Estimates of the Percentage of Adults at the Lowest Literacy Level for 1992 and 2003

Research and Development Report

US Department of Education
NCES 2009-482

National Assessment of Adult Literacy

Indirect County and State Estimates of the Percentage of Adults at the Lowest Literacy Level for 1992 and 2003

Research and Development Report

January 2009

Leyla Mohadjer
Graham Kalton
Tom Krenzke
Benmei Liu
Wendy Van de Kerckhove
Lin Li
Westat

Dan Sherman
Jennifer Dillman
American Institutes for Research

Jon Rao
Carleton University, Ottawa, Canada

Sheida White
Project Officer
National Center for Education Statistics

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Sue Betka
Acting Director

National Center for Education Statistics
Stuart Kerachsky
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006–5651.

January 2009

The NCES World Wide Web Home Page is http://nces.ed.gov/.
The NCES World Wide Web Electronic Catalog is http://nces.ed.gov/pubsearch/.

Suggested Citation
Mohadjer, L., Kalton, G., Krenzke, T., Liu, B., Van de Kerckhove, W., Li, L., Sherman, D., Dillman, J., Rao, J. and White, S. (2009). *National Assessment of Adult Literacy: Indirect County and State Estimates of the Percentage of Adults at the Lowest Level of Literacy for 1992 and 2003* (NCES 2009-482). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

For ordering information on this report, write to
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794–1398

or call toll free 1-877-4ED-PUBS or order online at http://www.edpubs.org/webstore/content/search.asp.

Content contact
Sheida White

Executive Summary

In 2003 the National Center for Education Statistics (NCES) conducted the National Assessment of Adult Literacy (NAAL) to measure the nation's English literacy skills, following up on the National Adult Literacy Survey (NALS), conducted in 1992. The 2003 NAAL interviewed over 18,500 adults (age 16 and older) across the country in private households. The overall sample comprised a core national sample supplemented by samples in six states that participated in the State Assessment of Adult Literacy (SAAL). The SAAL was designed to provide estimates of adult literacy levels for each of the participating states.[1] In a similar fashion, the 1992 NALS interviewed over 24,000 adults in private households, consisting of a core national sample supplemented by samples in the 11 states that participated in the State Adult Literacy Survey (SALS).[2,3]

The two surveys were designed to provide standard survey estimates—direct estimates—of literacy proficiency with adequate levels of precision for the target population for the nation as a whole, for major population subgroups (e.g., subgroups defined by region, level of educational attainment, and race/ethnicity) within the nation, and also for those states participating in the SAAL or SALS. However, based on the survey data alone, neither survey was designed to provide policymakers and educators estimates of the percentages of adults at the lowest literacy level at the state or county. Thus, NCES undertook a project to produce estimates of adults at the lowest literacy level for individual states and counties using statistical modeling approaches. These model-dependent estimates are called "indirect" estimates to distinguish them from standard or "direct" estimates that do not depend on the validity of a statistical model. The county and state indirect estimates were produced using small area estimation techniques that rely on survey data as well as data from other sources such as the decennial censuses for each of the two survey years.

This report describes the statistical methodology used to produce the model-dependent—indirect—estimates of the percentages of adults at the lowest literacy level for individual states and counties for 1992 and 2003. The county and state indirect estimates themselves are provided at the NAAL website http://nces.ed.gov/NAAL (the state indirect estimates are also provided in appendices to this report).

---

[1] The 2003 SAAL states were Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma.

[2] The 1992 SALS states were California, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, and Washington.

[3] In addition to the household samples, both surveys included samples of adults from federal and state prisons. The inmate samples did not contribute to the indirect county and state estimates presented in this report.

The NAAL and NALS produced direct estimates of Prose, Document, and Quantitative literacy[4], each reported on a 0 to 500 scale and on four performance levels: *Below Basic, Basic, Intermediate*, and *Proficient* based on this scale. The measure chosen for the indirect estimation is the percentage of adults lacking *Basic* prose literacy skills (*BPLS*). The literacy of adults who lack *BPLS* ranges from being unable to read and understand any written information in English to being able to locate easily identifiable information in short, commonplace prose text, but nothing more advanced.[5] It should be noted that adults who were not able to take the assessment because they were not able to communicate in English or Spanish (i.e. language barrier cases) are included in the indirect estimates and classified as lacking *BPLS* because they can be considered to be at the lowest level of English literacy. Users should note that the indirect estimates of the percentages lacking *BPLS* are not comparable to the percentages *Below Basic* in prose literacy in other NAAL or NALS published results because the latter excludes adults who were unable to take the assessment due to a language barrier.

The statistical model used to produce the indirect county estimates of the percentages of adults lacking *BPLS* was developed using the 2003 NAAL data; the same modeling approach was then applied to the 1992 NALS data. A Hierarchical Bayes (HB) model was adopted using a Markov Chain Monte Carlo (MCMC) method. The model was implemented using the WinBUGS software (Lunn et al. 2000). The key component of the approach was to develop a logit model (linear logistic regression model) to predict county percentages of adults lacking *BPLS* based on the survey data and a set of predictor variables that were available and measured consistently for all counties.

The process of model development involved the compilation of a large number of predictor variables that were known to be correlated with literacy from past analyses or hypothesized to be correlated with literacy (such as education, immigration, racial and ethnic minority status, age, employment status, occupation, urban/rural status, and poverty status). The list of candidate predictor variables was reduced to a manageable set based on bivariate analyses of the associations of these variables with the county direct estimates of the percentage lacking *BPLS* (for those counties for which direct estimates could be made). This set of candidate predictor variables was further reduced to a subset that was evaluated in the 2003 HB modeling. A similar subset was also considered for the 1992 model. The predominant source for the predictor variables that ended up in the final statistical model was the

[4] Prose literacy is the knowledge and skills needed to search, comprehend, and use continuous texts. Document literacy is the knowledge and skills needed to search, comprehend, and use non-continuous texts in various formats. Quantitative literacy is the knowledge and skills needed to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials. For more information on the three types of literacy, see http://nces.ed.gov/naal/literacytypes.asp.

[5] For more information about performance levels, see White and Dillow (2005) or see http://nces.ed.gov/naal/perf_levels.asp.

preceding Population Census (the 2000 Census for the 2003 model, and the 1990 Census for the 1992 model). Also, both models included predictor variables relating to education attainment, race/ethnicity, poverty status, indicators for census divisions, and state assessment indicators. There are some differences between the two models. For instance, extensive model testing resulted in including foreign-born and poverty status in the 2003 model only, while native English speaking status was used in the 1992 model only. For both years, 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.

A variety of methods was used to evaluate the fit of the HB models to the county direct estimates. The final models used to produce the county and state indirect estimates were insensitive to different model assumptions, and the measures of model fit indicated good fits to the data. The results from the measure of fit tests were similar for the NAAL and NALS models.

The precision of the indirect estimates of the county and state percentages of adults lacking *BPLS* depended heavily on the ability of the predictor variables in the model to predict these percentages. The critical importance of including variables that are effective predictors in the logit model is demonstrated by the fact that the NAAL collected data from 11 percent, and the NALS collected data from 13 percent of US counties. The indirect estimates produced for counties not in the samples therefore rely totally on the model predictions. The indirect estimates of counties that are included in the sample also relied heavily on the model predictions because their direct estimates were based on small samples and are generally imprecise. The median coefficient of variation of the direct estimates (i.e. the ratio of the standard error to the estimate) is 53 percent.

Although care was taken to select the sets of variables available that best predicted the county percentages of adults lacking *BPLS* in 1992 and 2003, and the sets did have a statistically significant relationship to the direct estimates, their predictive ability was limited, as reflected in the prediction error of the indirect estimates. Credible intervals have been computed to indicate the prediction error (i.e. levels of uncertainty) in the indirect estimates.[6] Users need to pay careful attention to the 95 percent credible interval bounds that are provided along with the indirect estimates to assess the range of uncertainty in the estimates. In general, the credible intervals tend to increase in size as the size of the point estimate increases.

---

[6] A credible interval is a posterior probability interval, used in Bayesian statistics (Bayes methods were used to create the small area models) for purposes similar to those of a confidence interval in frequentist statistics, with the exception that credible intervals are nonsymmetric around the estimate. A 95 percent credible interval for an estimate of the percentage of adults in a county lacking *Basic* prose literacy skills gives the range for which there is a probability of 0.95 that the interval contains the true percent lacking *BPLS*.

Overall, the levels of precision of the 1992 and 2003 HB model estimates for sample counties are comparable. The county estimates have median coefficients of variation (CV) of 33 percent for 2003 NAAL and 35 percent for 1992 NALS.[7] Thus, for example, for a county with an estimated 14 percent of adults lacking *BPLS* (approximately the national average for both years) and a CV of 35 percent, the 95 percent confidence interval (as an approximation to the credible interval) is roughly from 4 percent to 24 percent[8].

The state estimates are more precise, with median CVs of 14 and 15 for 2003 NAAL and 1992 NALS, respectively. For example, for a state with an estimated 14 percent lacking *BPLS* with a CV of 15 percent, the 95 percent confidence interval is from 10 to 18 percent.

Overall, the analysis of 1992 and 2003 results indicated that gains in precision were achieved in the estimates for SAAL and SALS states as a result of the larger sample size. Although the main purpose of the SAAL and SALS samples was to enable states to produce reliable direct estimates of literacy levels for all scales, at all levels, and for their major subgroups, the larger sample sizes in these states were also beneficial in producing more precise indirect estimates of the state percentages of adults lacking *BPLS.*

In addition to the need for county and state estimates of low literacy, policymakers and educators will often be interested in making comparisons between states and between counties. Credible intervals for the differences in the indirect estimates for pairs of states and counties (within states) in 2003 and for the differences in the indirect estimates for the same county or state between 1992 and 2003 have been computed and are available at the NAAL website (http://nces.ed.gov/naal/). Readers should keep in mind that the credible intervals for the differences in county indirect estimates are wide (a median width of 22 for estimates in the same state and year), which could be related to the limited ability to statistically detect differences in literacy levels based on 95 percent credible intervals. For example, while some differences can be detected between two counties within most states in NAAL, there are 7 states for which no significant differences between counties can be detected. Similarly, of the 3,100 comparisons made, 1 percent of the 1992 and 2003 county level differences are statistically detectable. At the state-level, 9 percent of apparent differences between 1992 and 2003 are statistically detectable.

---

[7] While the credible interval is the primary measure of precision, the CV provides a means to measure the variation relative to the point estimate. It is computed as the standard error divided by the point estimate.

[8] The CV is equal to the standard error divided by the point estimate. Therefore, the standard error for this example is equal to .35 * .14 = .049. Then the lower bound of a 95 percent confidence interval is computed as .14-1.96 * .049, which is equal to .044. The upper bound is computed as .14 + 1.96 * .049, which is equal to .236. Therefore, an approximate 95 percent confidence interval is from 4 percent to 24 percent.

As noted earlier, the model-based approach was used to create indirect estimates because there is no data source available that can provide reliable direct estimates of the percentage of adults at the lowest literacy level for all counties and states in the nation. As indicated above, the indirect estimates are not precise. However, they are offered as predictions that can be made from the national survey data. In the absence of any other literacy assessment data available for individual states and counties, the estimates provide a general picture of the status of literacy for all counties and states. Lacking these estimates, census variables highly correlated with literacy, such as educational attainment and poverty, have generally been used as proxy indicators of state and county literacy levels.

This page left intentionally blank.

Foreword

The Research and Development (R&D) series of reports at the National Center for Education Statistics has been initiated to

- Share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available;
- Share the results of studies that are, to some extent, on the "cutting edge" of methodological developments. Emerging analytical approaches and new computer software development often permit new and sometimes controversial analyses to be done. By participating in "frontier research," we hope to contribute to the resolution of issues and improved analysis; and
- Participate in discussions of emerging issues of interest to educational researchers, statisticians, and the Federal statistical community in general.

The common theme in all three goals is that these reports present results or discussions that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore, the techniques and inferences made from the data are subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done. Such responses should be directed to

Marilyn Seastrom
Chief Statistician
Statistical Standards Program
National Center for Education Statistics
1990 K Street NW
Washington, DC 20006-5651

This page left intentionally blank.

Acknowledgements

The authors gratefully acknowledge the many individuals who contributed to the preparation of this report.

First our thanks to the expert panel team, Partha Lahiri and Alan Zaslavsky, for their review of the technical aspects of the model development and selection process, and various aspects of the prediction approach, and we acknowledge their invaluable comments and suggestions that helped improve the estimates.

At the Education Social Science Institute (ESSI), we wish to thank Jaleh Soroui for managing and guiding the review process, Phuong Le for facilitating the review process, Shijie Chen for his technical review and comments on the small area estimation modeling aspects, and Enis Dogan, Jennifer Jeremias, Steven Osterlind, and Christian Vilenas for providing comments and recommendations that are reflected in this report.

At Westat, our thanks to Martha Berlin, the project manager, for her support and guidance. Finally, we extend our appreciation to Wen-Chau Haung, James Fan, and Eugene Brown for providing expert programming support for the intensive computer tasks unique to this estimation task.

This page left intentionally blank.

Contents

List of Appendixes

List of Tables

List of Appendix Tables

List of Figures

This page left intentionally blank.

1. Introduction

This report describes the statistical methodology used to produce U.S. county and state indirect estimates of the percentage of adults at the lowest literacy level based on survey data from the 2003 National Assessment of Adult Literacy (NAAL) and the 1992 National Adult Literacy Survey (NALS). The surveys are designed to measure the ability of adults to perform literacy tasks similar to those that they encounter in their daily lives. The 1992 NALS and 2003 NAAL assessments are on the same scale and are linked, that is, the literacy levels use the same framework and the literacy levels are comparable between the two assessments (White and Dillow 2005). Based on the survey data alone, neither survey was designed to provide policymakers and educators with estimates of the percentages of adults at the lowest literacy level for states and counties. Thus, NCES undertook a project to produce estimates of adults at the lowest literacy level for individual counties and states using statistical modeling approaches.

The main reason for including both the 1992 NALS estimates and 2003 NAAL estimates is to permit trend analysis. Another reason to provide the 1992 NALS estimates is because there are alternative 1992 NALS county estimates available on the web[9] that were not developed by NCES (that is, NCES had no input in their development) that have a relatively high degree of precision. The 1992 NALS indirect estimates given in the current report provide a more reasonable estimate of the precision (adequately captures sources of variance mainly due to the inclusion of random effects terms, as described in section 4.1) using a small area estimation methodology approved by NCES and similar to what is used in other government programs, like the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program.

In this document, the steps taken in the development of the statistical model and in the production and evaluation of the final estimates are described. The model development and evaluation were carried out using data from the 2003 survey[10]. Once a final model was developed for the 2003 NAAL survey, the same estimation method was applied to the 1992 NALS survey using similar variables (i.e., from the 2000 Decennial Census) to create estimates of literacy at the county and state levels[11].

---

[9] See http://www.casas.org/home/index.cfm?fusection=home.showContent&MapID=124.

[10] See http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008466 for more information on the 2003 NAAL survey.

[11] The main objective of this task was to produce model-based estimates for the 2003 NAAL and to replicate the same methodology for the 1992 NALS to arrive at comparable models. Therefore, the 1992 NALS modeling was not carried out at the same intensity as the 2003 models. More discussion is provided in chapter 7.

The 2003 NAAL was a large-scale survey of English literacy levels of adults aged 16 years and older residing in households in the United States. The NAAL, funded by the National Center for Education Statistics (NCES), followed the same procedures as those used in the 1992 NALS,[12] the first nationwide U.S. survey of adult literacy. Over 18,500 adults participated in the household component of the NAAL. This sample was made up of a basic national sample of adults supplemented by state samples in six states including approximately 5,800 adults that participated in the State Assessment of Adult Literacy (SAAL)[13]. In addition to the household component, approximately 1,200 inmates of federal and state prisons were assessed. The inmate sample did not contribute to the NAAL indirect estimates.

Each individual who participated in the NAAL provided demographic and other background information, and completed 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[14]. A set of booklets containing different sets of tasks was used, where each booklet contained less than a quarter of the tasks 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[15]: *Below Basic, Basic, Intermediate*, and *Proficient*. The NAAL reports provide results for the literacy levels of adults for each of the three scales separately. There were 3 percent of adults that were unable to complete a minimum number of simple literacy screening cases and were given an alternative assessment that were verbally asked in either English or Spanish, but all written materials were in English only. These 3 percent were included in the NAAL survey, and were included in the small area estimation modeling as well.

The approach used to collect data for the 2003 NAAL was similar to the approach used in the 1992 NALS. Over 24,600 adults participated in the household component of the NALS. The assessment was designed to produce national statistics to measure the literacy of the adult population in 1992, using a national sample of approximately 13,600 individuals. The national sample was supplemented by samples of about 1,000 individuals in the 11 states that participated in the State Adult

---

[12] A full description of the design of the 1992 NALS is available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2001457 or Kirsch et al (2000). The NAAL technical report (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008466) provides details on the design of the 2003 NAAL.

[13] The SAAL states were Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma.

[14] Prose literacy is the knowledge and skills needed to search, comprehend, and use continuous texts. Document literacy is the knowledge and skills needed to search, comprehend, and use non-continuous texts in various formats. Quantitative literacy is the knowledge and skills needed to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials. For more information on the three types of literacy, see http://nces.ed.gov/naal/literacytypes.asp.

[15] For more information on about the skills assumed at each literacy level, refer to White and Dillow (2005).

Literacy Survey (SALS).[16] Additionally, about 1,100 inmates of Federal and state prisons were given the assessments. The inmate sample was not used to develop the indirect estimates from NALS presented in this report. Several changes were made to the 1992 NALS data after their public release to improve their comparability with the 2003 data including new scales that measured literacy in terms of levels used in the 2003 NAAL: *Below Basic, Basic, Intermediate*, and *Proficient*.

The NAAL and NALS sample sizes are 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 and SALS surveys are able to produce 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 survey design does not support state-level estimation). Indeed, some states and most counties have no sample in the surveys.

Thus, NCES has used a statistical modeling (small area estimation) approach to produce model-dependent estimates of the percentages of adults in the lowest literacy 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 are derived directly from responses of individuals who live in an area included in the assessment. The indirect estimates are produced using small area estimation techniques that rely both on literacy estimates from other geographic areas included in the assessment and on other variables such as educational attainment that are available for all counties from data produced by other sources (such as the decennial Census). This approach uses sample information from all counties to "borrow strength" in producing the indirect estimates. By creating a model that predicts literacy levels for counties in the sample from the predictor variables, the model can then be used to make estimates for all counties and states. Rao (2003) and Jiang and Lahiri (2006) provide comprehensive overviews and comparisons of models and methods for small area estimation.

The choice of the percentage of those who lack *Basic* prose literacy skills (*BPLS*) for the small area estimation was made on the grounds that this measure reflects the magnitude of the adult household population at the lowest level of literacy (the prose scale measures the knowledge and skills needed to understand and use information from text). The literacy of adults who lack *BPLS* ranges from being unable to read and understand any written information to being able to locate easily identifiable information in short, commonplace prose text in English. For the indirect estimates, adults who were not able to communicate in English or Spanish and could not be tested are included since they can be

[16] The SALS states were California, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, and Washington.

considered to be at the lowest level of English literacy. These adults are known as "language barrier cases." The percentage of adults in the *Below Basic* group and those not able to take the assessment because of a language barrier is termed the percentage lacking *BPLS* in this report. Users should note that the indirect estimates of the percentages lacking *BPLS* are not comparable to the percentages *Below Basic* in prose literacy in other published results because the direct estimates of literacy levels exclude adults who are unable to take the assessments because of a language barrier.

A single area-level hierarchical regression model is used to predict the percentages of adults lacking *BPLS* in both states and counties. The model uses the direct county-level estimated percentages of adults lacking *BPLS* as the dependent variable in the model. County- and state-level variables that are related to adult literacy, obtained from the decennial censuses (1990 and 2000) and other reliable data sources (such as the American Community Survey[17] and the Behavioral Risk Factor Surveillance System[18]), are used as potential predictors of the dependent variable. The model also includes random state and county effects.

Hierarchical Bayesian (HB) estimation techniques with noninformative prior distributions are used to model the relationship between the predictor variables and the direct county estimates (the dependent variable). The posterior distributions[19] for the model parameters are used to produce the indirect estimates of the percentage of adults lacking *BPLS* for counties in the NAAL/NALS sample with direct estimates. Such counties are referred to as "sampled counties." The final model fitted to the sampled counties is then used to produce estimates for non-sampled counties using an HB approach. The state estimates are created by aggregating the county estimates, again using an HB approach.

It is important to take the prediction error in model-dependent indirect estimates into account in their interpretation. This error can be substantial. The NAAL and NALS indirect estimates are no exception. Users need to pay careful attention to the 95 percent credible interval bounds[20] that are provided for the NAAL and NALS indirect estimates to indicate the range of uncertainty in the estimates.

---

[17] Refer to http://www.census.gov/acs/www/ for more information on the American Community Survey.

[18] Refer to http://www.cdc.gov/brfss/ for more information on the Behavioral Risk Factor Surveillance System.

[19] The posterior distribution is the conditional probability distribution of the unobservable quantity, given the observed data. See, for example, Gelman (2004).

[20] A credible interval is a posterior probability interval, used in Bayesian statistics for purposes similar to those of a confidence interval in frequentist statistics, with the exception that credible intervals are non-symmetric around the estimate. A 95 percent credible interval for an estimate of the percentage of adults in a county lacking *Basic* prose literacy skills gives the range for which there is a probability of 0.95 that the interval contains the true percent lacking *BPLS*.

As mentioned earlier, the model development and evaluation was carried out using data from the 2003 survey. Once a final model was developed, the same estimation method was applied to the 1992 survey. Chapters 2 through 5 contain information about the NAAL and the development of the model using data from the NAAL. More specifically, chapter 2 of this report contains some background information on the NAAL, including the sample design and selection procedures, the definition of the percentage of adults lacking *BPLS*, and a description of the IRT approach used to produce direct estimates of these percentages.

Chapter 3 describes the numerous state- and county-level 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.

Chapter 4 describes the HB estimation technique used (Rao 2003) to create a single area-level model for producing the state and county-level estimates. It describes the explicit small area models used for the NAAL and the Markov Chain Monte Carlo (MCMC) approach used (See Gelman, et al (2004) and Robert and Casella (1999) for a description of the methodology) to obtain estimates of the model parameters. The chapter also 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 in this chapter, followed by a description of methods used to conduct comparisons between pairs of counties and states.

The small area modeling approach used for estimating the percentages of adults lacking *BPLS* assumes that the relative variances, or 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. Chapter 4 includes a description of the modeling approach used to smooth the estimated relvariances of the direct estimates.

Chapter 5 describes the details of the model fitting and testing. It describes the set of models chosen as the final candidates, and how the final model was selected. The section summarizes various steps taken to evaluate the model and the indirect estimates. It explains why benchmarking the county estimates to aggregated direct NAAL survey estimates was not employed.

Chapter 6 contains the small area estimation approach used for the analyses of the 1992 NALS. It includes some background information on the design and sample selection for the NALS, the predictor variables considered and used in modeling, and the evaluation of the model and the indirect estimates. Finally, chapter 7 provides a comparison of the 2003 NAAL and 1992 NALS models, and suggestions on how to compare the indirect estimates across the two surveys when examining trend.

2. The 2003 NAAL Survey[21]

Sponsored by the National Center for Education Statistics (NCES), the 2003 National Assessment of Adult Literacy (NAAL) was designed to measure the nation's adults' literacy skills. This chapter provides background information on the NAAL sample design and estimation practices and the use of NAAL data for the small area modeling. To begin, section 2.1 summarizes the NAAL sample design. The NAAL procedure for the estimation of proficiency in English literacy is explained in section 2.2, including implications for the small area modeling. The computation of direct county estimates of the percentages of adults lacking *Basic* prose literacy skills (*BPLS*), which was the initial step in producing indirect county and state estimates, is described in section 2.3.

2.1 The NAAL 2003 Sample Design

The NAAL 2003 household study was designed to be a nationally representative sample of persons in households or college dormitories who were 16 years of age or older (called "adults" below) at the time of interview, from the 50 states and the District of Columbia. The NAAL employed trained survey interviewers to conduct interviews with a sample of over 18,500 adults. Nested within the NAAL design were six state-level samples, with an aggregate sample size of about 5,800 adults. These state samples were designed to generate direct estimates for six participating states—Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma—called the State Assessment of Adult Literacy (SAAL). NAAL was also designed to provide high-precision national estimates for Blacks and Hispanics. To accomplish this, oversampling was carried out for these two subgroups in the national sample.

The NAAL sample was selected based on a four-stage sample design aimed at reducing the cost of interviewing and assessing respondents in their homes. The first stage of selection was of primary sampling units (PSUs). PSUs were defined to be counties or sets of counties with the following general characteristics:

- PSUs were required to have a minimum population of 15,000 persons.
- PSUs were required to be no wider than 100 miles in maximum point-to-point distance.

[21] Following authors contributed to this chapter: Tom Krenzke and Leyla Mohadjer, Westat.

- PSUs consisted of counties that were either all Metropolitan Statistical Area[22] (MSA) or non-MSA.

- PSUs were required to stay within state boundaries.

A total of 1,884 PSUs were formed, and 100 PSUs were selected with probability proportionate to size as the first-stage sample, with the estimated size equal to the year 2000 population. There were 16 certainty PSUs. The remaining 84 sampled PSUs were selected from a heavily stratified sample. An additional 74 PSUs were sampled for the SAAL states.[23]

The second stage of sampling was of segments. Segments were individual census blocks if they contained at least 60 households, or if not, combinations of adjacent blocks were formed within census tract boundaries to yield segments with at least 60 households. Segments were selected within sampled PSUs with a probability proportionate to size; the measure of size for a segment was a function of the number of year-round housing units within the segment. In NAAL, the Black and Hispanic populations were sampled at a higher rate than the remainder of the population to increase their sample size. This was accomplished by assigning a larger measure of size to high-minority segments in which Black and Hispanic adults accounted for 25 percent or more of the population.[24]. For SAAL, there was no oversampling. In total, 1,959 segments were sampled from the 100 PSUs, with an additional 861 sampled for SAAL.[25]

The third stage of selection was households within segments (a total of about 35,500 selected households in the combined NAAL and SAAL sample), and the final stage of selection was adults within households (one sampled adult for households with up to three adults, and two sampled adults for households with four or more adults). A total of about 23,500 persons were selected, resulting in about 18,500 persons who completed the background questionnaire. The data collection for the household sample was conducted from May 2003 through February 2004. In addition, approximately 1,200 inmates of federal and state prisons were assessed.

Interviewers, some of whom were bilingual in English and Spanish, visited households to select and interview adults. Each study participant was asked to answer questions about his or her

---

[22] As defined by the Office of Management and Budget, a Metropolitan Statistical Area is a core based statistical area with at least one urbanized area that has a population of at least 50,000. The Metropolitan Statistical Area comprises the central county or counties containing the core, plus adjacent outlying counties having a high degree of social and economic integration with the central county as measured through commuting.
[23] Of which 14 overlapped with the 84 national noncertainty PSUs.
[24] Nonminority households in these segments were deselected at a rate so that their sampling rate was equal to that of nonminority households in low minority segments.
[25] Of which two overlapped with the 1,959 segments selected for NAAL.

demographic characteristics, educational background, reading practices, and other areas related to literacy and then to respond to a series of diverse English literacy tasks included in the NAAL assessment.

2.2 Measuring Proficiency in English Literacy

The NAAL English literacy assessment included three components (1) prose literacy, (2) document literacy, and (3) quantitative literacy. The NAAL used a set of four categories: (1) *Below Basic,* (2) *Basic*, (3) *Intermediate*, and (4) *Proficient* to describe the literacy levels of the adult population in prose, document, and quantitative literacy. The proficiency scores ranged from 0 to 500, with those scoring at 210 or below in prose falling into the *Below Basic* literacy level. Section 2.2.1 provides background on the approach used to create proficiency estimates for respondents.[26] A small percentage (2 percent) of adults in the sample could not be tested because they were not able to communicate in English or Spanish (referred to as "language barrier cases"). The language barrier cases are described further in section 2.2.2.

2.2.1 Computation of Proficiency Literacy Estimates

A large number of tasks were administered in the NAAL assessment to ensure the survey covered a broad range of literacy tasks (tasks that simulated the demands adults encounter when they interact with written prose materials on a daily basis). However, to keep the testing time at a reasonable level, each participant was given a subset of the pool of literacy tasks using a matrix sample design in a way that ensured that each of the tasks was administered to a nationally representative sample of adults, with some core tasks being administered to all sampled adults. The NAAL cognitive test items for the literacy assessment tasks were all open-ended questions with one of four scores coded for each: (1) correct, (2) incorrect, (3) omitted, or (4) not reached.[27]

---

[26] The NAAL technical report (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008466) provides details on the psychometric properties and scaling of the 2003 NAAL assessment.

[27] The NAAL technical report (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008466) provides a description of the number of items included in the NAAL and the administration time. For the core/main assessment, each respondent was administered 7 core assessment items and approximately 25 main assessment items. The average administration time was 45.7 minutes for the core/main assessment.

Because different respondents took different sets of items that could be different in level of difficulty, it would be inappropriate to base the literacy estimates simply on the number of correct answers obtained. Therefore, large-scale assessments using matrix sampling rely on Item Response Theory (IRT) models (Birnbaum, 1968; Lord, 1980). The IRT model uses the item responses for each individual and regards the latent proficiency score as random. The NAAL technical report (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008466) provides details on the IRT modeling used for NAAL. The IRT modeling is implemented using the AM software package (American Institutes for Research and Cohen, J. 2006),[28] which relies on marginal maximum likelihood (MML) estimation.[29]

As mentioned above, each individual respondent is presented with only a relatively small sample of the literacy tasks, resulting in uncertainty in an individual's proficiency estimate. It is important to take into account the variances associated with the IRT estimates when assessing literacy statistics. Since the NAAL direct estimates are produced using IRT modeling, their variances not only reflected the sampling error, but also included a component associated with the IRT modeling.

To evaluate the implications on the small area modeling, the components of variance were examined for the national NAAL direct estimate of the percentage lacking *BPLS* and a selection of 15 NAAL direct county estimates. The AM software produces the following two variance components for the variance of a direct literacy proficiency estimate for a small area:

- variance of the posterior mean of the distribution of possible estimates of whether an individual in that area lacks *BPLS* (i.e. the component associated with IRT modeling), and
- mean of the posterior variance of the distribution of possible estimates of whether an individual in that area lacks *BPLS* (i.e. the component associated with sampling error).

For the national estimate, 38 percent of the total variance was attributable to the mean posterior variance. Examination of the percentage of total variance attributable to the mean posterior variance for the selected 15 NAAL county estimates did not indicate a relationship with sample size or magnitude of the direct variance. The percentages were, however, highly unstable, ranging across the 15 counties from 6 percent to 96 percent. These results did, however, show that the mean posterior variance

[28] AM is a statistical software package for analyzing large-scale assessment data from complex samples. Refer to http://am.air.org for more information.
[29] The following references provide more information on MML as offered in the AM software: Binder (1983), Bock and Aitkin (1982), Mislevy,et al (1992), Mislevy, Johnson, and Muraki (1992), and Mislevy and Sheehan (1987).

is a sizeable proportion of the total variance of the county estimates. This finding has implications for the modeling of the relvariances of the county estimates discussed in chapter 4.

2.2.2 Language Barrier Cases

Two percent of adults sampled for NAAL could not be tested because they could not communicate in English or Spanish[30]. These language barrier cases were included in the NAAL target population under the classification of Nonliterate in English. For the direct NAAL survey estimates in other published results, the language barrier cases did not contribute to the NAAL survey's literacy estimates. For the NAAL indirect estimates, however, the language barrier cases contributed to the estimates of the percentage lacking *BPLS* because they can be assumed to be at the lowest level of literacy. The language barrier cases were included in the small area estimation modeling by imputing a wrong response to the easiest core item in the prose assessment and then computing county-level direct estimates using the AM software.

In addition, one percent of adults could not be tested because of a mental disability that precluded conducting the NAAL interview. These adults do not contribute to direct NAAL survey estimates or the indirect estimates of the percentage lacking *BPLS*.

2.3 Direct County Estimates

As indicated in section 2.2.1, the AM software was used to generate direct estimates of literacy proficiency levels from the NAAL data based on the MML method. This methodology was applied to provide direct estimates of the percentages lacking *BPLS* for sampled counties for use in the small area modeling. The NAAL sample was restricted to the household sample; the sample of inmates was excluded. Section 2.3.1 provides information about the application of the AM software to produce these county-level direct estimates.

One aspect of the NAAL estimation approach is that direct estimates (e.g., literacy scores, or the percentage in a particular proficiency level) produced for subgroups are based on IRT models that are different from the model used for the aggregate group. If the group is made up of two or more subgroups,

---

[30] Spanish is mentioned since bilingual (English or Spanish) interviewers were able to assist the sample adults through the background questionnaire. If the participant could not answer a minimum number of simple literacy screening cases, they were given an alternative assessment where there was verbal assistance in either English or Spanish, but all written materials were in English only.

this implies the estimate for the total group cannot be generated from the subgroup estimates. Thus, strictly speaking, state estimates cannot be produced by combining the estimates for all the counties in the state. Section 2.3.2 discusses and illustrates this issue in the context of the NAAL direct and small area estimation.

For use in the small area modeling, the variances for the direct county estimates were estimated using the Taylor series approximation[31] in the AM software (see, for example, Wolter [1985] on the Taylor series approximation method of variance estimation). The direct variance estimates not only reflect the sampling error, but also measure the variances coming from the IRT estimation approach.

2.3.1 County-Level Direct Estimates of Percentages Lacking *Basic* Prose Literacy Skills

For use in the small area modeling, attempts were made to produce direct estimates of the percentages lacking *BPLS* for 324 of the 342 counties represented in the NAAL sample. Eighteen counties were excluded because they had fewer than five sampled adults. The percentage lacking *BPLS* was estimated separately for each county using the prose items in the assessment and with scores below 210 indicating a lack of *BPLS*. County-level direct variance estimates were produced by treating county as the variance stratum, and segment as the variance unit.

Direct estimates were obtained for 264 of the 324 counties with five or more sampled cases. Direct estimates were not obtained for 36 of the remaining 60 counties because they each had only one sampled segment, while at least two sampled segments were needed for variance estimation. For 24 counties direct estimates were not obtained because the estimation procedure failed to converge.[32]

As stated in section 1, the 264 counties with direct estimates will be referred to as sampled counties. The remaining 2,877 counties in the United States are referred to as non-sampled counties (even though some of them did have some sampled adults).

[31] For the Taylor series approximation for variance estimation, the first-order Taylor series linear approximation for the estimator is derived. The variance of this linearized estimate is then calculated as appropriate for the sample design, and used as an approximate variance estimate.

[32] Convergence was related to the number of segments and number of sampled persons in the county.

2.3.2 IRT Modeling: Aggregated Estimates of Subdomains Compared to Domain Estimates

A feature of the NAAL MML approach is that the weighted aggregate of subdomain direct estimates does not equal the direct estimate for the full domain. This feature implies that the aggregate of estimates for counties will not be equal to the direct estimate for the state, region, or nation, or any other domain of interest that comprises a combination of counties. Because this feature of the NAAL IRT modeling impacts the small area estimation approach, the magnitude of the discrepancy between these estimates was examined, and the results are as described below.

For each of the SAAL states, the direct estimate for the state and the weighted estimate derived from direct county estimates were computed. To produce a state estimate from aggregated county estimates, the counties with no direct estimates (see section 2.3.1 for an explanation of why some counties have no direct estimates) were combined with other counties to obtain direct estimates for the group. There were 8 such counties in Kentucky, 7 in Maryland, 7 in New York, 4 in Oklahoma, 11 in Missouri, and 1 in Massachusetts. Within all the states except Massachusetts, the counties with NAAL sample for which direct estimates were not obtained were grouped together to get one direct estimate for the group. The single county in Massachusetts with NAAL sample but no direct estimate was paired with another county that did have a county estimate to get a direct estimate for the pair. Then, using the predicted probabilities of lacking *BPLS* for sampled adults associated with the direct county estimates and the survey weights, state estimates were derived as combinations of county estimates. Table 2-1 compares these estimates with the direct state estimates. The table shows that in all cases the county-based estimates were larger than the direct state estimates, but the differences were small, within 95 percent confidence intervals of direct estimates for each of the six states. The implications of this finding for the production of state estimates from the small area model are discussed in chapter 5.

Table 2-1. Percentage lacking *Basic* prose literacy skills for direct state estimates and weighted aggregates from direct county estimates, by State Assessment of Adult Literacy states: 2003

State	Sample Size	Direct estimate	95 percent Confidence interval		County-based estimate	Difference
			Lower bound	Upper bound		
Kentucky	1,500	11.5	9.54	13.46	12.1	0.6
Massachusetts	1,100	10.7	7.96	13.44	11.7	1.0
Maryland	1,000	9.4	6.66	12.14	9.8	0.4
Missouri	1,000	7.1	5.14	9.06	7.5	0.4
New York	1,700	20.6	16.88	24.32	21.5	0.9
Oklahoma	1,300	12.5	9.36	15.64	12.6	0.1

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

3. 2003 NAAL Predictor Variables for the Small Area Estimation Modeling[33]

A key aspect of small area estimation modeling for the 2003 National Assessment of Adult Literacy (NAAL) was finding predictor variables that are measured consistently across all counties and states and that are effective predictors of the estimate of adults lacking *Basic* prose literacy skills (*BPLS*). The importance of identifying literacy-related predictor variables is magnified for NAAL, since units in the NAAL sample came from 11 percent of the counties in the United States. The remaining counties rely on a small area estimation prediction model (discussed in chapter 4) that uses direct NAAL survey estimates from counties containing NAAL sample, as well as data from sources other than the NAAL survey. In addition, the predictor variables help to improve the precision of estimates for counties that have NAAL sample.

To begin, a list of county and state level predictor variables was created from several sources, including 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). The predictor variables considered for this process are discussed in section 3.1.

Once the set of predictor variables was accumulated, a two-phase variable selection process was implemented. The result of this process is the set of predictor variables retained for the small area model. One objective in the selection process was to reduce the number of predictor variables to a level to ensure that the Markov Chain Monte Carlo iterations in the Hierarchical Bayes (HB) approach converge for all the model parameters (refer to section 4 for more discussion on convergence). Another motivating factor was to reduce the multicollinearity between the final predictor variables in the model. The selection process is described in section 3.2, and the final set of predictor variables is provided in section 3.3.[34]

3.1 County and State Predictor Variables

Given the importance of identifying predictor variables, a considerable effort was devoted to identifying reliable data sources and variables that are potential predictors of literacy. In total, over

---

[33] Following authors contributed to this chapter: Tom Krenzke, Graham Kalton, Leyla Mohadjer, Lin Li and Wendy Van de Kerckhove, Westat.

[34] The main objective of this task was to produce model-based estimates for the 2003 NAAL and to replicate the same methodology for the 1992 NALS to arrive at comparable models. Therefore, the 1992 NALS modeling was not carried out at the same intensity as the 2003 models.

100 variables across 20 major variable types (e.g., poverty, income, education, occupation, etc.) were obtained as potential predictors for the percentage lacking *BPLS*.

Appendix A provides details about the source, year, and level (state or county) of each variable considered for the small area model. The primary source was county-level data from the 2000 Census of Population. Summary File 3 (SF3) was used to extract county-level variables. The SF3 contains the Census "short form" items (items asked of all households) and includes information about age, gender, race, Hispanic or Latino origin, household relationship, and owner/renter status. The SF3 also contains the "long form" data coming from questions asked of about one-sixth of America's households. The questions include such topics as income, education, language spoken, housing structure, housing costs, and commuting. As shown in appendix A, in addition to the Census of Population, various other sources were used for obtaining county-level and state-level variables, for example, the Bureau of Economic Analysis (BEA) per capita personal income estimates for local areas, the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program, and the U.S. Department of Agriculture (USDA) Economic Research Service Rural-Urban Continuum Codes program.

Most of the variables are percentages of the county or state that fall into a specific category of the variable. For example, for country of birth, the following set of initial variables was formed:

- Percentage of foreign-born people who stayed in U.S. for 5 years or less;
- Percentage of foreign-born people who stayed in U.S. for 6 to 20 years;
- Percentage of foreign-born people who stayed in U.S. for 20 years or less; and
- Percentage of foreign-born people who stayed in U.S. for 21 years or more.

The selection process, as described below, focused on variables that were significant predictors of percentage lacking *BPLS*, regardless of whether the variables reflected the population distribution from 2000 or a later year. The 2003 U.S. Census Bureau intercensal estimates were not used since the estimated residency counts includes other populations outside the scope of the NAAL small area estimation population, including group quarters and institutions. It was therefore decided to use the Census 2000 numbers rather than making adjustments to the 2003 estimates at the county-level by race and gender. Population differences may be less of an issue at the state-level, however, in selecting predictors, state-level variables were considered only when the associated county-level variables (i.e., Census 2000 variables) were not available.

3.2 *County and State Predictor Variable Selection Process*

Variables listed within each variable type are generally highly correlated, as expected. For example, among the 264 sampled counties, for poverty variables, the percentage below the poverty line and the percentage below 150 percent of poverty are highly correlated ($r = 0.9$). In addition, several variables are highly correlated across variable types. For example, the percentage below 150 percent of poverty is highly correlated with median household income ($r = -0.9$), or percentage with less than a 9th grade degree ($r = 0.8$). The variable selection process was designed to address the issues relating to highly correlated predictor variables.

The process of selecting variables was conducted in two phases. In the first phase, the long list of county and state-level variables was reduced through (1) correlation and stepwise regression analyses between the predictor variables and the percentage lacking *BPLS*; (2) a review of sample design variables (i.e. variables used in sample selection) with impact on small area modeling; and (3) a review of variables known to be correlated with literacy from past analyses or hypothesized to be correlated with literacy. Once the lists of predictor variables were reduced, the second phase evaluated the variables using both empirical and Hierarchical Bayes models. The statistical testing mentioned in this chapter used the .05 level of significance.

Phase 1

The initial variable selection process excluded sampled counties with less than 50 observations to guard against unstable county estimates adversely affecting the identification of significant predictor variables. Another goal of phase 1 was to take precautions against multicollinearity effects on model results. To do so, in the search for variables that define the model predictors for the estimated percentage lacking *BPLS*, a stepwise regression analysis was processed between the logit of the percentage lacking *BPLS* and each of the county-level variables for each variable type separately. The regression procedures identified significant main effects for each variable type. Subsequently, the significant main effects for each variable type model were put into one model along with first-order interaction terms among the main effects. A stepwise regression was then processed to identify significant terms in the model. The model-fitting procedures in phase 1 of the variable selection process identified three county variables and one geographic variable for the final model — percent foreign-born in the country for 0-20 years, percentage with a high school education or less, percentage Black or Hispanic, and a census division indicator[35], respectively (refer to appendix A for the source and year of the variables). These variables showed statistically significant relationships with literacy. One (SAAL indicator) was not significant but was retained as a key variable to address the oversampling features of the sample design. The percentage Black or Hispanic variable was both a significant predictor and a key sample design variable (area segments with a high Black or Hispanic concentration were oversampled in NAAL). The five aforementioned variables were considered essential (i.e., core predictors) for the small area model.

At this point, state-level covariates were gathered and evaluated in an attempt to supplement the list of five core predictors (refer to appendix A for a list of state-level variables, and their source and year). State-level variables were not given any further consideration if conceptually similar county-level variables were also available. All other variables that were hypothesized as being related to literacy were downloaded, or key-entered. These variables are flagged in appendix A (table A-2). Once obtained, the variables entered a stepwise regression selection process along with the five aforementioned essential variables. Birth rate and financial aid received (percentage of full-time first-time students receiving any financial aid for their attendance at a Title IV institution), were found to be statistically significant, however financial aid received was dropped from consideration due to difficulty with convergence in subsequent test processing of Hierarchical Bayes models.

The set of variables identified as significant predictors in the correlation analysis and model-fitting process described above excluded some county variables that were hypothesized as correlates of literacy (e.g., percentage in poverty, percentage in service occupations, and percentage in agricultural occupations). Additional variables were added to the retained pool of predictor variables from phase 1. In addition, other state-level variables were given further consideration because of a moderate correlation with the direct estimate (e.g., violent crime rate, and adult education enrollment[36]). In general, a pairwise correlation was considered as having 'moderate' correlation if it had a correlation coefficient between .2 and .6. These variables were retained for phase 2 of the variable selection process. In total, 25 variables entered the phase 2 process. The list included the five core variables (percent foreign-born in the country for 0-20 years, percentage with a high school education or less, percentage Black or Hispanic, the census

[35] An indicator variable was created to identify a combination of census divisions. The indicator variable was equal to one if the county was in the New England, East North Central and or West North Central divisions, and equal to zero otherwise.

[36] Refer to appendix A for further details about the variables.

division indicator, and the SAAL state indicator), 8 state-level variables[37], and 12 other county-level variables[38].

The covariate selection process in phase 1 did not account for the sample design associated with various variable sources, if any, from which the variable was gathered. The phase 2 process, described below, addressed the sample design issue by bringing in the sample design variables (SAAL indicator, and Black or Hispanic variable), but did not address the sampling error in the long form items and other survey estimates. However, the final predictors were all from the Census 2000 and either not subject to sampling error (i.e., Census short form item relating to race/ethnicity) or subject to minimal sampling error (i.e., Census long form items relating to education attainment, poverty status, foreign born status).

Phase 2

The purpose of the phase 2 process was to evaluate the five core predictors under random effects models, and to determine if any other county-level or state-level variables (mentioned above) should be retained for further examination of the fit of the final small area model. The initial phase 2 process involved running mixed-effects models with the five core predictors, state and county random effects, and alternative sets of predictor variables, using empirical Bayes models. Empirical Bayes models were used initially because the processing time was much less than Hierarchical Bayes. Small sets of the additional county-level and state-level variables were systematically added to the list of five key variables during several runs. All statistically significant variables from the separate runs were pooled together in a final run. The significant variables and the sample design variables from the final run became the focus of the subsequent more exhaustive model building process via Hierarchical Bayes (refer to section 5.1 for a discussion of this process in the context of model diagnostics).

---

[37] The state-level variables retained for phase 2 included: birth rate, violent crime rate, infant mortality rate, health care coverage rate, graduation rate, percentage of civilian population that are veterans, percent of grandparents living with grandchildren and responsible for them, and household size.

[38] Besides the core variables, the other county-level variables retained for phase 2 included: percentage in poverty, percentage in service occupations, percentage that live less than 30 minutes from work, percentage that are home owners, percentage married, percentage who live in a different house since 1995, percentage that speak English not at all or not well, unemployment rate, percentage male, percentage with a employment disability, percentage divorced, and percentage that work in the county.

3.3 Predictor Variables in the Model

Chapter 5 contains the results of the extensive model evaluations carried out prior to selecting the final model. Table 3-1 provides the predictor variables retained in the final small area model. Table 3-2 shows the correlation matrix for these variables for the 264 sampled counties. The largest correlation coefficient, 0.8, is between the education and poverty variables. This is an acceptable level for a bivariate correlation between independent variables in a prediction model (a model solely used for prediction purposes). Furthermore, tolerance[39] levels, computed using ordinary least squares regressions, were also at an acceptable level, ranging from 0.3 to 0.9 across the predictor variables.

Table 3-1. List of predictor variables for the final small area model: 2003

Predictor	Level	Source
Percentage of the population who are foreign-born persons that stayed in the United States 0-20 years	County	2000 Census of Population
Percentage of persons age 25 and older with a high school education or less	County	2000 Census of Population
Percentage of the population who are Black or Hispanic	County	2000 Census of Population
Percentage of the population below the 150 percent poverty line	County	2000 Census of Population
Indicator variable identifying the New England and North Central census divisions	State	2000 Census of Population
Indicator variable identifying the State Assessment of Adult Literacy states	State	2003 National Assessment of Adult Literacy

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

[39] Tolerance is computed as 1 - R^2 for the ordinary least squares regression of that predictor variable on all the other predictor variables, ignoring the dependent variable. As a rule of thumb, if tolerance is less than .2, a problem with multicollinearity is indicated.

Table 3-2. Correlation coefficients among predictor variables for the final small area model: 2003

Variable	Percentage of persons age 25 and older with a high school education or less	Percentage of the population who are Black or Hispanic	Percentage of the population below the 150 percent poverty line	Indicator variable identifying the New England and North Central census divisions	Indicator variable identifying the State Assessment of Adult Literacy states
Percentage of the population who are foreign-born persons that stayed in the United States 0-20 years	-0.37	0.63	-0.10	-0.19	-0.23
Percentage of persons age 25 and older with a high school education or less		-0.14	0.75	-0.08	0.25
Percentage of the population who are Black or Hispanic			0.19	-0.28	-0.32
Percentage of the population below the 150 percent poverty line				-0.19	0.19
Indicator variable identifying the New England and North Central census divisions					-0.04

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

This page left intentionally blank.

4. 2003 NAAL Small Area Model Development and Prediction[40]

This chapter starts with a description of the Hierarchical Bayes (HB) model used to produce county and state indirect estimates of the percentages of adults lacking *BPLS* and the methods employed for fitting the small area model using the 2003 NAAL data and the WinBUGS software. Section 4.2 presents an account of the methods used to smooth estimates of the relative variances, or relvariances, of the county-level direct estimates for use in the HB models. Estimates of the model parameters for the final model are presented in section 4.3. Section 4.4 describes the methods used to produce the indirect estimates of the percentages of adults lacking *BPLS* for sampled counties, for non-sampled counties, and for states. The computation and interpretation of credible intervals for these estimates are described and discussed in section 4.5. Finally, methods for estimating credible intervals for differences between indirect estimates are presented in section 4.6.

4.1 Model for Indirect Estimates

A single HB model has been used to produce both county and state indirect estimates of the percentages of adults lacking *BPLS*. 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), and in You and Rao (2002).

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 lacking *BPLS* in county j in state i where $j = 1, \dots c_i$, and $i = 1, \dots m$. 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.

---

[40] Following authors contributed to this chapter: Leyla Mohadjer, Graham Kalton, Jon Rao, Benmei Liu, Tom Krenzke and Wendy Van de Kerckhove, Westat.

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 (less than 20 for 21 percent of sampled counties) and the values of θ_{ij} are also small. This assumption is, however, required for the assumed HB model, and follows Rao (2003) 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 hypothesized to be unstable because of small sample sizes. 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 (see section 4.3), again following the general approach in Rao (2003).

Linking model

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

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

where $logit(\theta_{ij}) = \log[\theta_{ij} / (1-\theta_{ij})]$, 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, ν_i is a state random effect ($\nu_i \overset{iid}{\sim} N(0, \sigma_\nu^2)$), and u_{ij} is a county random effect ($u_{ij} \overset{iid}{\sim} N(0, \sigma_u^2)$).

The following widely used prior distributions (Rao 2003) are assumed for the parameters on the right-hand side of the linking model:

Improper flat prior distribution, leading to a proper posterior distribution, on the vector of regression parameters β_k ;

$\sigma_v^2 \sim ING(0.001,\ 0.001)$, where *ING* denotes the inverse gamma distribution[41]; 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 (linear logistic regression model). If the linking model was linear, then the combination of the sampling and linking models would have resulted in the well-known small area level linear mixed model of Fay and Herriot (1979). However, a logit model provides a better fit than a linear model for the NAAL estimation since the variable of interest is the proportion lacking BPLS.[42] As a result, the HB approach is used to fit the unmatched NAAL model. Model fitting consists of producing posterior distributions for all the model parameters:

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

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

---

[41] Gelman (2006) considers the limitations of using the inverse gamma prior. However, he examines the case of small number of areas and variance component close to zero using a simple mean model with random effect. He shows that the posterior using inverse gamma (IG) with parameters (1,1) could be quite different from the IG using (0.01, 0.01). Note that the posterior mode is at zero. This result is not surprising under his scenario. In our case we have a large number of small areas and the posterior mode of the variance component is away from zero. In addition, extensive sensitivity analyses were conducted, and are described in chapter 5. These analyses found very little change in the HB estimate and posterior variance.

[42] You and Rao (2002) show that the customary log transformation approach (to arrive at matched models) could result in biased estimates and underestimation of variance. It may seem reasonable to take the log transformation of the direct estimate to obtain matched models. However, You and Rao (2002) showed that a customary log transformation approach on the original estimates could lead to estimation bias and underestimation of variance when county sample sizes are small.

Full conditional distributions

Let $\vec{p} = (p_{11},..., p_{1c_1};...; p_{m1},..., p_{mc_m})'$, $\vec{\theta} = (\theta_{11},...,\theta_{1c_1};...;\theta_{m1},...,\theta_{mc_m})'$, $\vec{v} = (v_1,...,v_m)'$. The full conditional distributions for all the model parameters of the hierarchical model defined by (1)~(7) are as follows:

$$\theta_{ij} \mid \vec{p},\vec{v},\beta,\sigma_v^2,\sigma_u^2 \propto \frac{1}{\sqrt{\psi_{ij}}\theta_{ij}(1-\theta_{ij})}\exp\left\{-\frac{\left(p_{ij}-\theta_{ij}\right)^2}{2\psi_{ij}}-\frac{\left(\log it(\theta_{ij})-x_{ij}^{'}\beta-v_i\right)^2}{2\sigma_u^2}\right\}, \quad \text{for} \quad 0<\theta_{ij}<1,$$

$j=1,...,c_i;\ i=1,...,m$;

$$v_i \mid \vec{p},\vec{\theta},\beta,\sigma_v^2,\sigma_u^2 \sim N\left[\frac{\sigma_v^2\sum_{j=1}^{c_i}\left\{\log it(\theta_{ij})-x_{ij}^{'}\beta\right\}}{c_i\sigma_v^2+\sigma_u^2},\frac{\sigma_v^2\sigma_u^2}{c_i\sigma_v^2+\sigma_u^2}\right], \text{for } v_i\in R,\ \ i=1,...,m;$$

$$\beta \mid \vec{p},\vec{\theta},\vec{v},\sigma_v^2,\sigma_u^2 \sim N\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{c_i}x_{ij}x_{ij}^{'}\right)^{-1}\left\{\sum_{i=1}^{m}\sum_{j=1}^{c_i}x_{ij}\left(\log it(\theta_{ij})-v_i\right)\right\},\ \ \sigma_u^2\left(\sum_{i=1}^{m}\sum_{j=1}^{c_i}x_{ij}x_{ij}^{'}\right)^{-1}\right]$$

for $\beta\in R$;

$$\sigma_u^2 \mid \vec{p},\vec{\theta},\vec{v},\beta,\sigma_v^2 \sim ING\left[a_1+\frac{1}{2}\sum_{i=1}^{m}c_i,\ \ b_1+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{c_i}\left(\log it(\theta_{ij})-x_{ij}^{'}\beta-v_i\right)^2\right], \text{for } \sigma_u^2\in R^+;$$

$$\sigma_v^2 \mid \vec{p},\vec{\theta},\vec{v},\beta,\sigma_u^2 \sim ING\left[a_2+\frac{1}{2}m,\ \ b_2+\frac{1}{2}\sum_{i=1}^{m}v_i^2\right], \text{for } \sigma_v^2\in R^+;$$

Note: The posterior distribution of θ_{ij} is $f(\theta_{ij} \mid \vec{p})$. It involves multi-dimensional integrals and it does not have a closed form. Also, one cannot generate directly from the first conditional but can generate directly from the remaining full conditionals. In WinBUGS, the model and a set of initial values for the model parameters is specified and samples are generated from the full conditionals based on the given model using the Metropolis-Hastings rejection algorithm.

4.2 Smoothing the Direct Relative Variances

As indicated earlier, the HB model for estimating the percentage lacking *BPLS* assumes that the relvariances (φ_{ij}^2) of the direct county estimates are known, whereas in practice they are unknown.

Since the direct estimates of these relvariances are subject to substantial sampling error, the true relvariances have also been predicted using a modeling approach. A requirement of this modeling is that the predicted relvariances should not depend directly on the county-level direct estimates or variance estimates. An important feature of the development of the model for predicting the relvariances is that approximate values will suffice since the values of the relvariances affect the estimates of the percentage lacking *BPLS* in only a minor way. Their main impact is in stabilizing the widths of the credible intervals.

Since the relvariance of a direct county estimate depends on the value of the county's percentage lacking *BPLS*, a two-step approach was developed to produce model-dependent estimates of the relvariances. In step 1, the proportions lacking *BPLS* θ_{ij} were predicted from a simple regression model relating the direct estimates of p_{ij} to predictor variables selected from those listed in chapter 3. The set of variables in the step 1 model does not need to be the same as the set of final predictors for the small area model because the objective was the smoothing of relvariances, and it was one of the first stages of the small area estimation process, which preceded the model selection process. In step 2, the resulting predicted proportions from step 1 were used in a generalized variance function (GVF) model to smooth the relvariance estimates. The predicted values of the relvariances $\tilde{\varphi}_{ij}^2$ for the county proportions of adults lacking *BPLS* were then treated as known relvariances in the HB model. The following paragraphs provide details on the smoothing process.

As with the HB model, the logit of the proportion lacking *BPLS* was used as the dependent variable in the regression model.[43] Predictor variables were selected for step 1 using a stepwise selection method. A robust regression M-estimation approach using SAS Proc RobustReg was used to arrive at the predicted values of p_{ij}. 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 square root was applied as an *ad hoc* method of approximating weighting by residual variance.[44] Outliers (2 percent) were identified based on the default bisquare function, also referred to as Tukey's biweight. The bisquare function is described in detail in the discussion on the RobustReg procedure in the SAS/STAT user's manual (SAS Institute 2003). All the outliers were downweighted and none were set to zero.

[43] A model with the natural logarithm of the proportion lacking *BPLS* was also examined, but the logit model provided a better fit to the data.

[44] Three other modeling procedures were also examined: ordinary least squares, weighted least squares using the square root of the sample size as the weight, and weighted least squares using the sample size as the weight. In these three models, outliers were given a weight of zero. The predicted values of the proportions lacking *BPLS* from all the weighted least squares models were highly intercorrelated (i.e. a correlation of close to 1.0). The correlation of the predicted values from the unweighted ordinary least squares approach with those from each of the weighted approaches was 0.9.

The final model has the form:

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

where p_{ij} = the proportion lacking *BPLS*;

Z_{ij1} = the percentage of persons age 25+ with a high school education or less;

Z_{ij2} = the percentage of Blacks and Hispanics;

Z_{ij3} = the percentage of foreign-born persons who stayed in the United States 0-20 years;

Z_{ij4} = an indicator for New England and North Central census divisions; and

e_{ij} = the error term.

Table 4-1 presents the estimated parameters of the model processed among the 264 sampled counties, which had an R^2 value of .4. Although the emphasis is on the predicted values of the dependent variable, the parameter estimates are provided to show the magnitude of the relationship with the dependent variable and the direction of the relationship, while controlling for the effects of the other variables in the model. As a check, after the predictions were done using the 2003 NAAL HB model for sampled counties, the predicted values from the HB model were compared to the predicted values from step 1 of the relvariance smoothing process. The correlation coefficient between the two sets of predicted values was .9.

Table 4-1. Parameter estimates for the first step of the variance smoothing process for the county-level direct estimates of the proportion lacking *Basic* prose literacy skills: 2003

Parameter	Estimate	Standard error	95 percent confidence limits		Chi-square	p-value
γ_0	-3.7	0.23	-4.13	-3.23	257	<.001
γ_1	2.5	0.40	1.68	3.26	38	<.001
γ_2	1.1	0.32	0.45	1.71	11	<.001
γ_3	4.5	0.88	2.82	6.26	27	<.001
γ_4	-0.4	0.11	-0.59	-0.17	13	<.001

SOURCE: U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

In step 2, the predicted values of the proportions lacking *BPLS* from the regression model in equation (3) were used as predictor variables in the model to smooth the relvariance estimates. To make the model linear in the parameters, a robust weighted least squares log-log model, where the weight was the square root of the degrees of freedom for the direct variance estimate, was used. The robust regression approach is the same as the approach used in step 1.[45] The rationale for the use of the square root of degrees of freedom as weights in the GVF model is the same as that for the use of the square root of sample size in the regression model for predicting the proportion lacking *BPLS* (that is, the direct estimates of relvariances were weighted in the GVF regression by a measure that was related to their precisions). This ad hoc weighting scheme downplayed the less precise relvariances. The model has the form:

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

where φ_{ij}^2 = the relvariance of the proportion lacking *Basic* prose literacy skills;

$\tilde{p}_{ij}$ = the predicted proportion from equation (3);

n_{ij} = the sample size; and

e_{ij} = the error term.

This model draws on a sampling error model, where relvariances are functions of $\tilde{p}_{ij}$, $(1-\tilde{p}_{ij})$, and n_{ij}. However, in the current situation the relvariances also include the variance associated within the IRT modeling (see section 2.2.1 on the contribution of the IRT variance component). The model does not therefore have a solid theoretical basis.

The outliers (3 percent) were identified based on the bisquare function in Proc RobustReg. All the outliers were downweighted and none were set to zero. Table 4-2 contains the parameter estimates for the robust GVF regression processed on the 264 sampled counties. The model had an R^2 value of .4.

[45]Three other approaches were also considered: weighted least squares using the degrees of freedom as the weight, weighted least squares using the square root of degrees of freedom as the weight, and a design effect approach. In these approaches, outliers were given a weight of zero. The predicted values of the proportions lacking *BPLS* from all models were highly intercorrelated (over .8).

Table 4-2. Parameter estimates for the second step of the variance smoothing process for county -level relvariances: 2003

Parameter	Estimate	Standard error	95 percent confidence limits		Chi-square	p-value
η_0	0.2	1.01	-1.81	2.15	#	.863
η_1	-1.0	0.33	-1.6	-0.33	9	.003
η_2	0.2	1.58	-2.87	3.34	#	.883
η_3	-0.9	0.09	-1.08	-0.74	107	<.001

Rounds to zero

SOURCE: U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

The predicted values of the relvariances for the county proportions of adults lacking *BPLS* were computed based on the GVF regression model in equation (4), and these predicted values were then treated as known relvariances in the HB model.

4.3 Model Fitting

Model selection started with a preliminary comparison of different models with alternative sets of predictor variables, as described in chapter 3. A selected set of models was chosen to go through an extensive evaluation process. This resulted in a model with the six predictor variables listed in section 3.3. The evaluation process is described in chapter 5. This chapter describes the procedures employed to fit the final model with these six variables.

Model fitting was carried out using a Markov Chain Monte Carlo (MCMC) method. The WinBUGS software (Lunn et al. 2000), version 1.4, 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")[46] 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 $\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u}, \sigma_v^2, and\ \sigma_u^2$, 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

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

$\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u}$ 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 (SAS Institute 2003). The distributions of $\boldsymbol{\beta}, \boldsymbol{v}$, and $\boldsymbol{u}$, were assumed to be approximately normal. The MLE variances were varied by 10 percent (i.e., three levels of variances were used: 1) using the variances as is, 2) subtracting 10 percent, and 3) adding 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. The initial values $\boldsymbol{\beta}^{(0)}, \sigma_v^{2(0)}, \textit{and } \sigma_u^{2(0)}$ for each run of the final model are shown in table 4-3.

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, and 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 determined to have been reached. The iterations up to this point (called the burn-in period) were discarded.

Table 4-3. Initial parameter values for the Metropolis-Hastings algorithm, by run: 2003

	Run		
Parameters	1	2	3
Intercept	-3.8	-3.8	-4.1
Percentage of the population who are foreign-born persons that stayed in the United States 0-20 years	3.0	3.0	4.9
Percentage of persons age 25 and older with a high school education or less	2.1	3.0	3.2
Percentage of the population who are Black or Hispanic	1.8	1.0	1.8
Percentage of the population below the 150 percent poverty line	0.4	1.0	1.5
Indicator variable identifying the New England and North Central census divisions	-0.4	-0.5	-0.4
Indicator variable identifying the State Assessment of Adult Literacy states	#	0.1	-0.2
Variance of county random effect	0.1	0.1	0.1
Variance of state random effect	#	#	#

Rounds to zero.

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

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, a total of 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.

4.3.1 The Final HB Model

The results of the final HB model, processed on the 264 sampled counties, are shown in table 4-4 for the parameters $\boldsymbol{\beta}, \sigma_v^2$ and σ_u^2. Although the credible intervals for the SAAL indicator and the poverty variable include zero, these predictor variables were retained in the model. Extensive evaluations of the resulting estimates for the sampled and nonsampled counties showed that the inclusion of these variables in the model provided a better fit and improved the predicted values of θ_{ij} for nonsampled counties.

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, shown in the last column of table 4-4, 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. The Gelman-Rubin statistic R (Brooks and Gelman 1998) compares the ratio of the pooled chain variance to the within chain variance. 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. The values of $\hat{R}$ for $\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\theta}$ (not shown in table 4-4) are also all near 1. 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.

Table 4-4. Regression coefficients and variances of random effects for the final HB model: 2003

Parameters	HB mean	HB standard deviation	Median	95 percent credible interval Lower bound	Upper bound	$\hat{R}$
Intercept	-3.6	0.22	-3.6	-4.03	-3.18	1.0
Percentage of the population who are foreign-born persons that stayed in the United States 0-20 years	4.5	0.75	4.5	3.02	5.98	1.0
Percentage of persons age 25 and older with a high school education or less	2.2	0.56	2.2	1.12	3.28	1.0
Percentage of the population who are Black or Hispanic	1.0	0.32	1.0	0.39	1.66	1.0
Percentage of the population below the 150 percent poverty line	0.6	0.83	0.6	-0.95	2.31	1.0
Indicator variable identifying the New England and North Central census divisions	-0.4	0.11	-0.4	-0.58	-0.14	1.0
Indicator variable identifying the State Assessment of Adult Literacy states	-0.1	0.10	-0.1	-0.32	0.09	1.0
Variance of county random effect	0.1	0.03	0.1	0.06	0.19	1.0
Variance of state random effect	#	0.02	#	#	0.06	1.0

Rounds to zero

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

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. The WinBUGS program is provided here:

```
model {

	#N observations

		for (i in 1:N) {

			P[i] ~ dnorm(theta[i], D[i])
			D[i] <- 1/(RELVAR[i]*theta[i]*theta[i])
			logit(theta[i])<- beta1+beta2*X1+beta3*X2+beta4*X3+beta5*X5+beta6*X8

					+beta7*X9+v[state2[i]]+c[i]

			 c[i]~dnorm(0, sigma_c2)
				}
```

```
        # M states
for (j in 1:M) {
        v[j] ~ dnorm(0,sigma_v2)
                }

# Priors

beta1 ~ dflat()
beta2 ~ dflat()
beta3 ~ dflat()
beta4 ~ dflat()
beta5 ~ dflat()
beta6 ~ dflat()
beta7 ~ dflat()



sigma_c2 ~ dgamma(0.001, 0.001)
sigma_v2 ~ dgamma(0.001, 0.001)

var_c <- 1/sigma_c2
var_s <- 1/sigma_v2


      }
```

4.4 Predicted Values for Counties and States

As mentioned above, estimates for the parameters $\theta_{ij}^{(b)}$, $\boldsymbol{\beta}^{(\boldsymbol{b})}$, $u_{ij}^{(b)}$, $v_j^{(b)}$, $\sigma_u^{2(b)}$, and $\sigma_v^{2(b)}$ for $b = 1$, ..., 27,000 MCMC samples, were produced through the WinBUGS software for sampled counties. Once the final model was processed and the model parameters estimated, the next step was to estimate the percentage lacking *BPLS* for sampled counties, nonsampled counties, and for states. The prediction process for sampled and nonsampled counties is described in sections 4.4.1 and 4.4.2, respectively. The process for making state-level estimates is described in section 4.4.3.

4.4.1 Indirect Estimates for Sampled Counties

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

$$\hat{\theta}_{ij}^{HB} = \frac{\sum_{b=1}^{27,000} \theta_{ij}^{(b)}}{27,000} \tag{5}$$

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

$$logit(\theta_{ij}^{(b)}) = \boldsymbol{x'_{ij}}\,\boldsymbol{\beta}^{(b)} + v_i^{(b)} + u_{ij}^{(b)} \tag{6}$$

4.4.2 Indirect Estimates for Non-Sampled Counties

For sampled counties, estimates of all the components on the right hand side of equation (6) were 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. To simulate the MCMC procedure, in cases where a component was not available, it was drawn at random from the appropriate normal distribution. Thus, following Rao (2003), $u_{ij}^{(b)}$ was drawn from $N(0,\sigma_u^{2(b)})$ and, when necessary, $v_i^{(b)}$ was drawn from $N(0,\sigma_v^{2(b)})$.

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

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

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

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

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 (5).

4.4.3 Indirect Estimates for States

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.

Because county populations of the household residents 16 years or older were not available for 2003, the weight for each county was estimated using available data from the U.S. Census Bureau to create initial county estimates for the National Assessment of Adult Literacy (NAAL). The 2003 estimated residency counts from the U.S. Census Bureau include other populations outside the scope of the NAAL small area estimation population, including group quarters and institutions. Therefore the 2003 estimated residency counts for ages 16 and older were adjusted by the ratio of Census 2000 counts for persons within households to total population.

Then initial county population estimates were calibrated to the sum of the final NAAL sampling weights for the State Assessment of Adult Literacy (SAAL) states and the remainder of each census region to improve consistency between indirect and direct estimates.

4.5 Measures of Precision for the Indirect Estimates

The primary measure of precision reported for each NAAL state or county indirect estimate is its credible interval, described in section 4.5.1. An alternative measure of uncertainty is the coefficient of variation (CV), discussed in section 4.5.2. An assessment of the precision of the indirect estimates using both measures is provided in section 4.5.3.

4.5.1 Credible Intervals

A credible interval is a posterior probability interval, used in Bayesian statistics for purposes similar to those of a confidence interval in frequentist statistics.[47] A 95 percent credible interval is any interval with a probability under the posterior distribution of .95. For example, a statement such as "following the model result, a 95 percent credible interval for the HB estimate for θ is 7 percent to 21 percent" means that the posterior probability that θ lies in the interval from 7 percent to 21 percent is .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.

4.5.2 Coefficient of Variation

The coefficient of variation (CV) of the HB estimate for county *j* in state *i* is computed as

$$CV_{ij} = \frac{\sqrt{Var(\hat{\theta}_{ij}^{HB})}}{\hat{\theta}_{ij}^{HB}} \tag{9}$$

where the posterior variance $Var(\hat{\theta}_{ij}^{HB})$ is computed as

$$Var(\hat{\theta}_{ij}^{HB}) = \frac{\sum_{b=1}^{27,000} (\hat{\theta}_{ij}^{(b)} - \hat{\theta}_{ij}^{HB})^2}{27,000-1} \tag{10}$$

Similarly, for states, the CV is computed as

$$CV_{i} = \frac{\sqrt{Var(\hat{\theta}_{i}^{HB})}}{\hat{\theta}_{i}^{HB}} \tag{11}$$

[47] Frequentist statistics is an approach to statistics that defines probability in terms of the frequency of occurrence in a series of trials.

where the posterior variance is computed as

$$Var(\hat{\theta}_i^{HB}) = \frac{\sum_{b=1}^{27,000} (\hat{\theta}_i^{(b)} - \hat{\theta}_i^{HB})^2}{27,000 - 1} \tag{12}$$

4.5.3 Assessment of Precision Measures

Appendix B contains the final state indirect estimates and credible intervals. The final county indirect estimates and credible intervals are provided at the NAAL website (http://nces.ed.gov/NAAL). In general, the credible intervals tend to increase in size as the size of the point estimate increases. This can be seen in figure B-1 in Appendix B. Table 4-5 summarizes the distributions of the widths (the difference between the upper bound and the lower bound) of the credible intervals as well as the coefficients of variation (CVs) for the 3,141 counties in the US.

Table 4-5. Distribution of credible interval widths and coefficients of variation for indirect county and state estimates: 2003

Statistic	Percentile 20	Percentile 40	Percentile 60	Percentile 80	Median
County estimates					
95 percent credible interval width (percent)	10.7	13.0	16.1	20.3	14.5
Coefficient of variation (percent)	30.3	32.1	34.2	35.4	33.0
Sampled county estimates					
95 percent credible interval width (percent)	8.5	10.5	12.9	14.7	11.9
Coefficient of variation (percent)	22.4	26.4	29.2	33.0	27.7
Nonsampled county estimates					
95 percent credible interval width (percent)	10.9	13.3	16.5	20.7	14.9
Coefficient of variation (percent)	30.8	32.4	34.5	35.5	33.3
State estimates					
95 percent credible interval width (percent)	4.8	5.3	6.4	7.6	5.8
Coefficient of variation (percent)	10.6	13.1	15.2	18.1	14.0

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Overall, the county estimates are less precise than the state estimates. For example, the median credible interval width for county estimates is 15 percent (i.e., percentage points), while the median is 6 percent for state estimates. The table also shows that the median credible interval width is 12 percent for counties with NAAL sample cases and 15 percent for counties without NAAL sample.

The CVs for the indirect county estimates are on the order of 30 percent or more. Half of the 3,141 counties have a CV of more than 33 percent. Estimates with CVs of this magnitude are highly imprecise. It is important for the users of these county estimates to recognize this fact. While the state estimates are more precise, with a median CV of 14 percent, it is still important for users to consider the credible interval along with the indirect estimate.

Table 4-6 displays the credible intervals widths and CVs for the SAAL states and provides a summary for the non-SAAL states. The credible interval widths and CVs indicate that higher precision was achieved in the state estimates from the SAAL states compared to the non-SAAL states. The sample sizes for the SAAL states ranged from 900 to 1,600, whereas the sample sizes for the non-SAAL states with sampled counties ranged from 80 to 1,500. The CVs of the indirect state estimates for all the SAAL states are less than the 20th percentile of the CVs for the non-SAAL states. The table also shows that the CVs for the indirect estimates for the SAAL states appear to be smaller than the CVs for the direct estimates. For instance, the CV for Maryland's indirect estimate is 10 percent compared to 15 percent for its direct estimate. Apart from Kentucky, which had a larger SAAL sample size than other SAAL states, the precision of the state estimates was much improved by the modeling process. Although the main purpose of the SAAL samples was to provide states the ability to produce reliable direct estimates of literacy levels for all scales, at all levels, and for their major subgroups, their larger sample sizes were also beneficial in producing more precise indirect estimates for SAAL states. More comparisons between direct estimates and aggregates of indirect county estimates can be found in section 5.2.

Table 4-6. Credible interval widths and coefficients of variation of indirect and direct estimates, by State Assessment of Adult Literacy (SAAL) and non-SAAL States: 2003

State	Credible interval width of the indirect estimate (percent)	Coefficient of variation of the indirect estimate (percent)	Coefficient of variation of the direct estimate (percent)
SAAL state			
Kentucky	4.0	8.4	8.7
Maryland	4.6	10.5	14.9
Massachusetts	3.8	9.9	13.1
Missouri	3.2	11.0	14.1
New York	5.3	6.1	9.2
Oklahoma	4.1	8.6	12.8
Non-SAAL states			
20th percentile	4.8	11.5	†
40th percentile	5.7	13.8	†
60th percentile	6.6	16.2	†
80th percentile	7.8	18.4	†
Median	6.0	14.7	†

† Not applicable.

NOTE: The calculations were done on unrounded numbers.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

4.6 *Comparisons of Indirect Estimates*

The MCMC procedures were extended to provide credible intervals for the differences between any pair of counties or states for 2003 NAAL. For each MCMC sample, the quantity $(\hat{\theta}_{ij}^{HB} - \hat{\theta}_{i'j'}^{HB})$ was computed and the credible interval for the difference $(\hat{\theta}_{ij}^{(b)} - \hat{\theta}_{i'j'}^{(b)})$ was 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 for 2003 NAAL. Credible intervals for differences between counties in different states have to be approximated by other means (an approximation is provided below). Likewise, credible intervals need to be approximated in order to compare indirect estimates for 2003 NAAL and 1992 NALS for single counties or states, and to do pairwise comparisons of counties and states for 1992 NALS.

For 2003 NAAL, the credible intervals for the differences between pairs of states and between pairs of counties within the same state are made available to users at the NAAL website (http://nces.ed.gov/naal/) via a web tool. In general the credible intervals for the differences between two county indirect estimates are large; the median width is 22.. While some differences can be detected between two counties within most states, there are 7 states for which the credible intervals for the apparent differences include 0 for all comparisons.

Based on analyses of the between-state county comparisons, the following approximate methods are suggested for determining whether the 95 percent credible interval for the difference between the indirect estimates for two counties that are in different states contains 0:

1. If the credible interval for county *j* does not overlap with the credible interval for county *j'*, then one can conclude that the credible interval of the difference does not contain 0. For example if the credible interval for one county is from 6 percent to 12 percent, and 13 to 21 percent for another county, then the credible interval of the difference will not include 0.

2. If the credible interval for county *j* is fully nested within the credible interval for county *j'*, then the credible interval for the difference will contain 0. For example, if one county has a credible interval of from 6 to 18 percent, and another county has a credible interval from 7 to 17 percent, then the credible interval of the difference will include 0.

3. If the credible intervals between two counties partially overlap (e.g., the credible interval is from 6 to 18 for one county and from 12 to 24 for another county), the following conservative approach can be used to help determine whether the credible interval for the difference contains 0.

Approximate the standard error of the difference between the indirect estimates for the two counties, $\hat{D}^{HB}_{ij,i'j'} = \hat{\theta}^{HB}_{ij} - \hat{\theta}^{HB}_{i'j'}$ by the following:

$$SE_{\hat{D}^{HB}_{ij,i'j'}} = \sqrt{\left(\frac{CRIWIDTH_{ij}}{4}\right)^2 + \left(\frac{CRIWIDTH_{i'j'}}{4}\right)^2} \tag{13}$$

where *CRIWIDTH* represents the credible interval width. Then approximate the credible interval by $\hat{D}^{HB}_{ij,i'j'} \pm 2SE_{\hat{D}^{HB}_{ij,i'j'}}$.

This procedure was compared with the exact procedure (using the MCMC samples) for a subset of pairs of indirect estimates for counties in different states. Among the 9,000 pairwise differences computed, when the credible interval from the exact procedure contained 0, there was just one case using the approximation that did not contain 0. When the credible interval from the exact procedure did not contain 0, there were 73 percent that had a credible interval from the approximate procedure that also did not contain 0. The approximate procedure is thus conservative in the sense that it sometimes indicates that the credible interval contains 0 when it does not. In those cases where the results differed, 81 percent of the credible intervals from the exact procedure were less than a percentage point from 0. Attempts to develop an alternative approximation showed no improvement.

Using the above approach, credible intervals can be approximated for the differences between pairs of states and between pairs of counties within the same state for the 1992 NALS. Likewise, the approximation in equation (13) can also be applied when comparing indirect estimates for a single county or single state across the two survey years (the 2003 NAAL and the 1992 NALS). Discussed in more detail in section 7.2, the approximate method of creating credible intervals described above has been used to create approximate credible intervals of the differences between the 1992 and 2003 county and state indirect estimates. These credible intervals are available at the NAAL website via a web tool similar to the one created for the 2003 estimates, as described above.

5. 2003 NAAL Small Area Model Evaluation48

Several approaches were employed in evaluating the 2003 NAAL final Hierarchical Bayes (HB) model and the resulting predicted values for the state and county percentages of adults lacking *Basic* prose literacy skills (*BPLS*). First, section 5.1 compares estimates for the final model and several alternative models and evaluates measures of fit. Section 5.2 compares aggregated indirect county estimates to direct estimates for selected geographic domains. Section 5.3 provides a summary of the model evaluation process.

5.1 Evaluation of Alternative Models and Assessing the Fit of the Final Model

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 predictor 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.

Alternative Prior Distributions

The following variants of the prior distributions were examined. The variations mentioned below are typical of those used in general practice to examine how robust the model is to its assumptions.

- changing the noninformative flat prior distributions for the regression coefficients β to informative normal priors with mean 0 and variances on the order of 10^6;
- 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 are near 1.0,

[48] Following authors contributed to this chapter: Leyla Mohadjer, Graham Kalton, Jon Rao, Tom Krenzke, Benmei Liu and Wendy Van de Kerckhove, Westat.

indicating that the final estimates are not sensitive to the choice of the prior distributions. Regarding the priors on the regression coefficients, these are not surprising results since a normal distribution with large variance is similar to a uniform distribution. Given that all priors are non-informative, it is not surprising that the results are fairly similar, and the posterior distributions will be essentially determined by the data.

Alternative Predictor Variables

Over 15 models, with alternative sets of predictor variables, were compared in the model selection process. As mentioned in chapter 3, some models were selected for further evaluation using the extensive HB approach. This section compares nine models that were selected for the HB approach—one of which was the final model selected based on the measures described below. All of the models contained a core set of five variables that the earlier analyses reported in chapter 3 had shown to be important. These variables are reproduced in the upper section of table 5-1. Each model also included some of the additional variables listed in the lower section of the table, with different combinations of additional variables in different models. The additional variables were introduced into the various 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. Furthermore, three versions (continuous, square root of the percentage, and a dichotomous recode) of the percentage in farming/fishing/forestry were created to examine the impact of this type of variable on the model fit and prediction. The additional variables included in the various alternative models are indicated in table 5-2.

Table 5-1. List of predictor variables for select alternative models, including their label, source, year, and level: NAAL 2003

Predictor variables	Label	Source	Year	Level
Core predictor variables				
Percentage of the population who are foreign-born persons who stayed in the U.S. 0-20 years	Foreign born	2000 Census of Population	2000	County
Percentage of persons age 25 and older with a high school education or less	Education	2000 Census of Population	2000	County
Percentage of the population who are Black or Hispanic	Black or Hispanic	2000 Census of Population	2000	County
Indicator variable identifying the New England and North Central census divisions[1]	Census division	2000 Census of Population	2000	State
Indicator variable identifying the State Assessment of Adult Literacy (SAAL) states	SAAL state	NAAL	2003	State
Additional predictor variables				
Percentage of the population below the 150 percent poverty line	Poverty	2000 Census of Population	2000	County
Percentage of the population in service occupations	Service occupations	2000 Census of Population	2000	County
Percentage of the population in farming/ fishing/ forestry occupations	Farming/fishing/ forestry	2000 Census of Population	2000	County
Percentage of women 15 to 50 years old who had a birth in the past 12 months	Birth rate	American Community Survey	2003	State
Violent crime rate per 100,000 population	Violent crime rate	Federal Bureau of Investigation	2003	State

[1] An indicator variable was created to identify a combination of census divisions. The indicator variable was equal to 1 if the county was in the New England, East North Central and or West North Central divisions, and equal to zero otherwise.

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3. U.S. Department of Commerce, Census Bureau, American Community Survey (2003). U.S. Federal Bureau of Investigation, Crime in the United States (2003). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Table 5-2 also provides the correlation coefficients between the indirect estimates from the final model and the indirect estimates from each of the alternative models for the sampled counties. As shown in the table, these correlation coefficients were near 1.0 for sampled counties and for nonsampled counties, for all eight alternative models. Thus all the models produced similar predicted values.

The deviance information criterion (*DIC*) (Spiegelhalter et. al. 2002) was used to compare the fit of the alternative models. This measure, which is reported by WinBUGS, is a measure of goodness of fit that takes account of the number of parameters in the model (like an adjusted R^2). The DIC measure is a function of the deviance and the effective number of parameters. The deviance, $D(y,\theta)$, measures how well the model fits the data, and is defined as $D(y,\theta) = -2 \log p(y|\theta)$, where $p(y|\theta)$ is the likelihood function. The posterior mean of the deviance is denoted $\overline{D}$. The *DIC* measure is adjusted to account for the estimated effective number of parameters (*pD*). The *pD* is the posterior mean of the deviance minus the deviance of the posterior means. The DIC measure is computed as, $DIC = \overline{D} + pD$. 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 10 (BUGS 2004). This rule is analogous to the one for the Akaike Information Criterion (AIC) used for logistic regression models (Burnham and Anderson, 2004). The last column in table 5-2 shows the *DIC* results for model A is -526, and for the other models ranges from -532 to -537. Based on the *DIC* criterion model A was ruled out, leaving a choice to be made between the other models.

Since the *DIC* measure could not definitively identify one model as the best fit, other criteria factored into the decision. Models B, C, E, and H, which involved the percentage or square root of the percentage in farming/fishing/forestry were excluded because of the high level of extrapolation needed for some nonsampled counties that had much larger values for this percentage than the maximum observed percent in the sample data. This problem was avoided by the use of the dichotomous farming/fishing/forestry variable in model G, but the indirect estimates for several nonsampled counties from this model were found to be extremely different from those produced by other models, raising concerns about the reliability of these estimates. Model D—which contained the percentage in poverty and the state birth rate—and the final model—which contained the percentage in poverty—were slight improvements over model F, which contained only the core set of variables. Lastly, there was no difference in the *DIC* between the final model and model D. After extensive examination of these variables and evaluation of their impact on the final estimates, it was decided to keep the county-level poverty variable in the final model, and drop the state-level birth rate variable, since the additional state-level birth rate variable did not add any reduction to the *DIC* value.

Table 5-2. Predictor variables in the alternative models, correlation coefficients between the indirect estimates from the final model and the other models, and the deviance information criterion (DIC), by model: 2003

Model	Additional predictor variables included in the HB model[1]							Correlation of estimates with final model		DIC
	Poverty	Service occupations	Farming/ fishing/ forestry (continuous)	Farming/ fishing/ forestry (dichotomous)	Farming/ fishing/ forestry (square root)	Birth rate	Violent crime rate	Sampled counties	Non-sampled counties	
Final	x							1.00	1.00	-537
A						x	x	1.00	.99	-526
B	x	x	x			x		1.00	.96	-532
C	x		x					1.00	.98	-534
D	x					x		1.00	1.00	-537
E			x					1.00	.98	-536
F								1.00	1.00	-536
G	x			x				1.00	.97	-533
H	x				x			1.00	.98	-537

[1] In addition to the core variables that were in all the models: Percentage foreign-born, percentage with a high school education or less, percentage Black or Hispanic, census division indicator, and SAAL state indicator. The variable labels are shortened. For full description of the variables, refer to chapter 3

NOTE: An ‘x’ denotes that the model includes the additional predictor variable.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

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 are near 1.0 for sampled counties and nonsampled counties. It was therefore decided not to include the county weight as a predictor variable in the final model. The final model is discussed in section 4.1, and the list of variables used in the final model is given in table 4-2.

Measures of Model Fit

Once the final model was selected, the following three measures were computed on the 264 sampled counties to assess the goodness of fit.

- A 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 indirect county estimates. The posterior predictive *p* value is the proportion of the 27,000 Markov Chain Monte Carlo (MCMC) samples that had a smaller simulated discrepancy measure (as opposed to the direct discrepancy measure) and should be close to .5 if the model fits the data well. The predictive p-value approach is a widely used approach for model evaluation, however, it has limitations in that it can induce unnatural behavior. This is because the data are used twice: once to fit the model, and once again to assess the fit of the model. See Rao (2003) and Bayarri and Berger (2000) for more discussion. Since the distribution of the estimates of the percentage lacking *BPLS* deviate somewhat from a normal distribution and since the estimates were less than 10 percent for several counties, 7 percent of the simulated estimates were negative and were therefore excluded. After these exclusions, the *p* value for the final model was equal to .61, indicating a good fit to the data.

- A county-level measure computed as the proportion of the 27,000 MCMC samples that had a smaller simulated value (as opposed to direct estimates). These proportions are expected to vary markedly across the counties. However, there should be a small number of counties with values close to 0 or 1. Across the counties, the proportion ranged from .05 to .98, with a global average of .53. Among the 264 sampled counties, 262 (99 percent) of the county-level values fell within the range of 0.05 and 0.95. There were two extreme value counties, with proportions of .05 and .98. Based on this measure, the model fit is very good.

- A 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. Values of this measure between, say, -1.96 and 1.96 and a global average of around 0 may be considered to be reasonable. The values obtained ranged from -2.09 to 1.39, with a global average of -0.04. Among the 264 sampled counties, 263 (over 99 percent) of the county-level values fell within the range of -1.96 and 1.96. The one extreme county with the value of -2.09 was one of the two extreme values identified by the previous measure. Overall, this measure also supports the model.

5.2 *Comparison of Direct Estimates and Aggregates of Indirect County Estimates*

A 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. Because of the Item Response Theory (IRT) scaling methods used (as discussed in section 2.3.2), the direct estimate for a domain is not the same as a combination of the county direct estimates for that domain. However, as shown in table 2-1, the differences between these two forms of domain direct estimates for SAAL states are one percentage point or less. The comparisons reported in this section assume the differences are also small for the domains examined.

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 (as described in section 4.4.3). Also, for each domain, direct estimates were computed from the sample data using the NAAL IRT approach via the AM software. Table 5-3 shows the resulting estimates. The direct and indirect estimates of the percentage lacking *BPLS* for the nation are close, differing by 0.1 percentage points. By region, the differences tend to be somewhat larger (between 0.4 to 1.5 percentage points). However, in general, the differences are under 1 percentage point. The table also shows the associated standard errors for the direct estimates from which 95 percent confidence intervals can be constructed to provide a range of values likely to contain the true value of the percent lacking *BPLS*. The table does not show measures of uncertainty surrounding the indirect estimates, however, since the aggregated indirect estimates always fall within the 95 percent confidence intervals for the direct estimates, the general conclusion is that there is no detectable difference between the indirect and direct estimates shown in the table.

An issue related to these comparisons is whether to benchmark the indirect estimates to conform to direct estimates for certain large domains. Benchmarking can be 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.

Table 5-3. Comparison of aggregated indirect county estimates and direct estimates for percentage lacking *Basic* prose literacy skills, by subgroup: 2003

Subgroup (Source: Year) [1]	Indirect estimate: Number of counties	Indirect estimate: Weighted estimate	Direct estimate: Sample size	Direct estimate: Estimate	Direct estimate: Standard error	Percentage point difference	Relative difference (percent)[2]
Overall	3,100	14.6	18,400	14.5	0.60	0.1	0.6
Census region (CENSUS: 2000)							
Northeast	220	15.6	3,700	15.1	0.81	0.5	3.7
Midwest	1,100	8.9	3,600	8.5	1.03	0.4	4.6
South	1,400	15.9	8,200	17.4	1.34	-1.5	-8.8
West	450	17.6	2,900	16.2	1.17	1.5	9.0
Variables used in the model:							
Percentage Black or Hispanics (CENSUS: 2000)							
< 10.6	1,900	9.2	6,100	9.8	0.92	-0.6	-5.8
10.6 -- 28.8	640	11.2	6,100	12.1	1.04	-0.9	-7.3
≥ 28.8	620	22.5	6,100	22.6	1.18	-0.1	-0.6
Percentage with high school education or less (CENSUS: 2000)							
< 43.5	360	11.3	6,100	11.4	0.78	#	-0.1
43.5 -- 54.1	810	15.6	6,200	17.4	1.21	-1.8	-10.4
≥ 54.1	2,000	17.4	6,100	15.5	1.25	1.9	12.4
Percentage foreign-born (CENSUS: 2000)							
< 2.5	2,300	11.0	6,100	11.5	1.25	-0.6	-4.6
2.5 -- 7.9	640	11.0	6,100	10.8	0.77	0.1	1.4
≥ 7.9	200	21.2	6,100	21.3	1.40	-0.1	-0.4
Percentage below 150 percent poverty line (CENSUS: 2000)							
< 17.2	580	10.4	6,000	11.1	0.89	-0.6	-5.7
17.2 -- 23.4	880	13.8	6,200	14.5	0.91	-0.7	-4.5
≥ 23.4	1,700	21.0	6,200	20.3	1.62	0.7	3.2

See notes at end of table.

Table 5-3. Comparison of aggregated indirect county estimates and direct estimates for percentage lacking *Basic* prose literacy skills, by subgroup: 2003—Continued

Subgroup (Source: Year)[1]	Indirect estimate: Number of counties	Indirect estimate: Weighted estimate	Direct estimate: Sample size	Direct estimate: Estimate	Direct estimate: Standard error	Percentage point difference	Relative difference (percent)[2]
State Assessment of Adult Literacy (SAAL) indicator (NAAL: 2003)							
Non-SAAL states	2,700	14.5	10800	14.6	0.72	-0.1	-0.7
SAAL states	410	15.5	7,600	14.4	0.75	1.1	7.5
Census division (CENSUS: 2000)							
New England, East North Central, West North Central	1,100	8.8	4,900	8.4	0.80	0.4	5.1
Others	2,000	16.9	13,500	17.3	0.76	-0.4	-2.4
Variables not used in the model:							
Beale Codes (USDA: 2003)							
Non-metro area	2,100	13.1	3,500	12.9	1.56	0.2	1.5
Metro area of >=1 million population	410	16.3	9,900	15.9	0.91	0.4	2.4
Metro area of <1 million population	680	12.6	5,100	13.2	1.49	-0.6	-4.8
Estimated target population size (NAAL: 2003)							
<114,725	2,800	12.3	6,100	12.3	1.29	0.1	0.2
114,725 -- 580,780	310	11.9	6,200	12.4	1.03	-0.5	-3.9
>=580,780	70	19.4	6,100	19.1	1.08	0.2	1.2
Median housing value of specified owner-occupied housing units (ACS: 2003)							
< $108,600	1,300	14.2	5,700	13.2	1.40	1.0	7.9
$108,600 -- $186,000	1,500	12.5	6,200	14.1	1.16	-1.7	-11.7
>=$186,000	310	18.1	6,500	16.3	0.89	1.7	10.6
Percentage in service occupation (CENSUS: 2000)							
< 13.9	840	11.9	6,000	11.8	1.00	0.1	1.0
13.9 -- 15.6	820	14.7	6,100	13.7	0.96	0.9	6.7
>= 15.6	1,500	17.2	6,300	18.9	1.46	-1.7	-9.1

See notes at end of table.

Table 5-3. Comparison of aggregated indirect county estimates and direct estimates for percentage lacking *Basic* prose literacy skills, by subgroup: 2003—Continued

Subgroup (Source: Year)[1]	Indirect estimate		Direct estimate			Percentage point difference	Relative difference (percent)[2]
	Number of counties	Weighted estimate	Sample size	Estimate	Standard error		
Percentage that commute less than 30 minutes to work (CENSUS: 2000)							
< 58.4	550	19.7	6,100	18.6	1.06	1.2	5.6
58.4 -- 70.7	940	13.1	6,200	13.4	1.07	-0.3	-2.2
>= 70.7	1,700	11.9	6,100	11.8	0.96	0.1	0.7
Percentage of population 5 and over that speak other language and speak English not at all or not well (CENSUS: 2000)							
< 1.1	2,100	10.5	6,000	10.8	1.09	-0.2	-2.3
1.1 -- 4.0	700	10.9	6,200	11.2	0.97	-0.3	-2.4
>= 4.0	320	22.0	6,200	21.9	1.21	0.1	0.4
Average household size (CENSUS: 2000)							
<2.5	1,600	11.9	6,200	13.4	1.32	-1.6	-11.7
2.5 -- 2.6	710	13.6	6,400	12.7	1.11	1.0	7.6
>=2.6	870	18.0	5,800	17.5	0.98	0.5	2.7
Percentage in farming/fishing/forestry occupation (CENSUS: 2000)							
< 0.2	150	15.3	6,200	15.4	0.88	-0.1	-0.4
0.2 -- 0.6	540	13.9	6,100	13.7	0.82	0.3	1.9
>= 0.6	2,500	14.9	6,100	14.8	1.61	0.1	0.2
Percentage with health care coverage (BRFSS: 2003)							
< 83.3	1,100	16.3	6,200	17.8	1.70	-1.5	-8.3
83.3 -- 86.9	750	17.6	6,300	16.0	0.96	1.6	9.8
>= 86.9	1,300	9.7	5,900	10.3	0.93	-0.7	-6.5
Infant mortality rate (NCHS: 2002)							
< 6.4	720	17.0	6,000	15.4	1.06	1.6	10.4
6.4 -- 7.6	1,150	14.5	6,200	15.2	1.48	-0.7	-4.4
>= 7.6	1,300	12.1	6,100	13.2	1.14	-1.1	-8.3

See notes at end of table.

Table 5-3. Comparison of aggregated indirect county estimates and direct estimates for percentage lacking *Basic* prose literacy skills, by subgroup: 2003—Continued

Subgroup (Source: Year) [1]	Indirect estimate		Direct estimate			Percentage point difference	Relative difference (percent)[2]
	Number of counties	Weighted estimate	Sample size	Estimate	Standard error		
Average graduation rate for students (IPEDS: 2003)							
< 51.5	1,300	11.8	6,200	12.2	1.35	-0.4	-3.5
51.5 -- 58.7	1,100	17.3	5,000	16.5	1.04	0.8	5.0
>= 58.7	700	13.6	7,100	14.7	1.46	-1.0	-7.0
Average percentage for students receiving financial aid (IPEDS: 2003)							
< 75.0	560	16.2	6,400	16.4	1.29	-0.2	-1.0
75.0 -- 80.9	1,200	14.4	5,900	14.4	1.01	0.0	0.2
>= 80.9	1,400	12.2	6,100	12.5	1.43	-0.3	-2.3
Gross state product in current dollars (BEA: 2003)							
<198.0 million	1,500	11.5	6,600	10.9	1.12	0.7	6.0
198.0 million -- 499.7 million	1,100	11.3	6,100	12.2	1.06	-0.9	-7.3
>=499.7 million	540	20.3	5,700	21.0	1.37	-0.7	-3.3
Census Division (CENSUS: 2000)							
New England	70	8.7	1,400	8.2	0.74	0.5	6.8
Middle Atlantic	150	18.1	2,400	18.6	1.06	-0.5	-2.9
East North Central	440	9.6	1,900	9.6	1.63	0.1	0.8
West North Central	620	7.0	1,700	6.5	0.85	0.6	8.8
South Atlantic	590	15.7	3,500	17.9	1.96	-2.3	-12.6
East South Central	360	13.8	2,000	17.8	4.63	-4.0	-22.4
West South Central	470	17.4	2,700	16.3	1.72	1.1	6.7
Mountain	280	12.1	800	10.7	1.10	1.4	13.5
Pacific	170	19.9	2,000	18.7	1.55	1.2	6.3

See notes at end of table.

Table 5-3. Comparison of aggregated indirect county estimates and direct estimates for percentage lacking *Basic* prose literacy skills, by subgroup: 2003—Continued

Subgroup (Source: Year) [1]	Indirect estimate: Number of counties	Indirect estimate: Weighted estimate	Direct estimate: Sample Size	Direct estimate: Estimate	Direct estimate: Standard error	Percentage point difference	Relative difference (percent)[2]
State Assessment of Adult Literacy state (NAAL: 2003)							
Kentucky	120	12.2	1,500	11.4	1.00	0.7	6.1
Maryland	20	11.2	1,000	9.4	1.37	1.8	19.5
Massachusetts	10	9.9	1,000	10.7	1.43	-0.8	-7.2
Missouri	120	7.5	1,000	7.1	1.03	0.3	4.5
New York	60	22.1	1,700	20.6	1.86	1.5	7.1
Oklahoma	80	12.3	1,300	12.5	1.62	-0.3	-2.2

Rounds to zero.

[1] The following variables are state-level variables: median housing value of specified owner-occupied housing units, percentage with health care coverage, infant mortality rate, average graduation rate for students, average percentage for students receiving financial aid, and gross state product in current dollars. All other variables, with the exception of the SAAL state indicator, SAAL state, census region and census division, are county-level variables.

[2] The relative difference is computed as the difference divided by the direct estimate. Differences when conducting the relative difference using numbers shown in the table are due to rounding. The calculations were done on unrounded numbers.

NOTE: The acronyms for the data sources are: CENSUS = Summary File 3 from Census 2000, USDA = United States Department of Agriculture, ACS = American Community Survey, IPEDS = The Integrated Postsecondary Education Data System, NAAL = National Assessment of Adult Literacy, BRFSS = Behavioral Risk Factors Surveillance System, BEA = Bureau of Economic Analysis, NCHS = National Center for Health Statistics. The table does not show a measure of uncertainty surrounding the indirect estimates, however, since the aggregated indirect estimates always fall within the 95 percent confidence intervals for the direct estimates, the general conclusion is that there is no detectable difference between the indirect and direct estimates shown in the table.

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3; U.S. Department of Commerce, Census Bureau, American Community Survey (2003); U.S. Department of Agriculture, Economic Research Service (2000); Centers for Disease Control Behavioral Risk Factor Surveillance System (2003); National Center for Health Statistics Vital Statistics of the United States (2002); Bureau of Economic Analysis Survey of Current Business (2005); U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Integrated Postsecondary Education Data System (2003); U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

5.3 Conclusion for the Model Evaluation

Various techniques were used to evaluate the fit of the 2003 NAAL HB model to the observed data. First, alternative models were constructed using different prior distributions and different sets of predictor 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 all support the model used in creating the NAAL county- and state-level indirect estimates. However, it needs to be recognized that the resultant indirect estimates are imprecise (see section 4.5.3) and have an associated credible interval to give users an indication of the magnitude of the uncertainty surrounding the estimate. This situation arises because of the lack of powerful predictor variables, consistently measured and available for all counties, for use in model construction. Overall, as described in section 4.5.3, the state indirect estimates are more precise than county indirect estimates, and the SAAL indirect estimates are more precise than non SAAL indirect estimates. The small area estimation approach was used to create indirect estimates because there is no data source available that can provide reliable direct estimates of the percentage of adults at the lowest literacy level for all counties and states in the nation. As discussed in section 1, they provide a general picture of the literacy status for all counties and states.

6. 1992 NALS Small Area Estimation[49]

This chapter applies the small area estimation methodology described in chapter 4 to the National Adult Literacy Survey (NALS) data to estimate the percentage lacking *Basic* prose literacy skills (*BPLS*) at the state and county levels in 1992. The main reason for including both the 1992 NALS estimates and 2003 NAAL estimates is to permit trend analysis. Another reason to provide the 1992 NALS estimates is because there are alternative 1992 NALS county estimates available on the web[50] that were not developed by NCES (that is, NCES had no input in their development) that have a relatively high degree of precision. The 1992 NALS indirect estimates given in the current report provide a more reasonable estimate of the precision (adequately captures sources of variance mainly due to the inclusion of random effects terms, as described in section 4.1) using a small area estimation methodology approved by NCES and similar to what is used in other government programs, like the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program.

This analysis, which applies the methodology used to fit and evaluate statistical models for the 2003 data, was undertaken after the final model specifications for the 2003 data had been developed. The small area estimation model developed for the 2003 NAAL data was used with the 1992 NALS data, although alternative variables were considered in order to provide a better fit to the 1992 data. The analysis of the NALS data provides estimates for the percentages of adults lacking *BPLS* for states and counties in 1992 that can be compared to those obtained from the 2003 data. Comparisons of the 1992 and 2003 estimates are presented in chapter 7.

6.1 The 1992 NALS Survey

The 1992 NALS was a survey of the levels of English literacy of adults aged 16 and over residing in households in the United States. The survey, which was funded by the National Center for Education Statistics, was designed to produce national statistics to measure the literacy of the adult population, using a core sample of approximately 13,600 individuals. The core national sample was supplemented by samples of about 1,000 adults in the 11 states that participated in the State Adult

---

[49] Following authors contributed to this chapter: Dan Sherman and Jennifer Dillman, American Institutes for Research.

[50] See http://www.casas.org/home/index.cfm?fusection=home.showContent&MapID=124.

Literacy Survey (SALS).[51] Additionally, 1,100 inmates of Federal and state prisons were given the literacy assessments; however, this inmate sample was not used to develop the estimates from NALS presented in this report.

The approaches used to collect data for the NALS and NAAL were similar. Individuals were asked to provide demographic and other background information, and were asked to complete a series of literacy tasks that were used to develop estimates of prose, document, and quantitative literacy. Several changes were made to the 1992 data after their public release to improve their comparability with the 2003 data, including proficiency levels that measured literacy according to the levels used in the 2003 NAAL: *Below Basic, Basic, Intermediate*, and *Proficient*. Following the procedure used for the small area analyses of the 2003 data, individuals who could not be tested because they were unable to communicate in English or Spanish were included in the 1992 small area analyses.

The NALS and SALS data were collected using a four-stage sample design summarized below. While the 1992 national and state household samples were drawn using the same sampling strategy, the sample designs differed in two ways: Blacks and Hispanics were oversampled only in the national sample, and the target population for the national sample consisted of all adults age 16 or older while the target population for the state sample was limited to adults aged 16 to 64. Blacks and Hispanics were oversampled in the national sample in order to provide reliable statistics for these domains. The national sample was also designed to produce reliable statistics for the adult population and for persons aged 65 or older.

The four sampling stages for the national and state samples were: (1) the selection of primary sampling units (PSUs) consisting of counties or groups of counties, (2) the selection of segments consisting of census blocks or groups of blocks within sampled PSUs, (3) the selection of households within sampled segments, and (4) the selection of age-eligible individuals within sampled households. The sample selection steps in 1992 NALS were similar to ones described in section 2.1 for 2003 NAAL. For more details about the 1992 NALS sample design, refer to *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (Kirsch et al. 2000).

As indicated earlier, as part of the production of indirect estimates, changes were made to the 1992 measurement scales to enable valid comparisons to be made with the 2003 scales. Several items

[51] The SALS states were California, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, and Washington.

were recategorized from the prose to document literacy scales. In addition, several dichotomous items were rescored using a partial credit model. To accommodate these changes, the 1992 data were recalibrated to provide item characteristic parameters comparable to those obtained from the 2003 data. Data from the test blocks that were used in both the 1992 and 2003 assessments were pooled for this rescaling; 6 out of 13 blocks used in the 2003 assessment were also used in 1992 assessment. Because of the rescaling, results using 1992 data may differ slightly from the results produced for the original public release of the 1992 data.

6.2 Direct County Estimates

The first step in the estimation process was to compute direct estimates of literacy proficiency for individual counties included in the 1992 sample by applying the method of marginal maximum likelihood using AM software (as described in section 2.2.1). The NALS sample collected data from 409 counties (out of 3,141 in the nation), with the numbers of sampled individuals in the sampled counties ranging from 3 to 776 around a median sample size of 41. Direct estimates of the percentages of adults lacking *BPLS* were obtained for 368 of these counties, which represented 98 percent of the individuals sampled for the NALS. Convergence was not reached in the remaining 31 counties because of small sample size and low number of segments. In this chapter, the 368 counties for which these direct estimates could be computed will henceforth be referred to as "sample counties."

The direct county estimates for the 368 counties included both estimates of (1) the percentage lacking *BPLS* and (2) the standard errors of these estimates. The standard error estimates were produced using a Taylor series approximation. Given the relatively small sizes in most county samples, the direct estimates were generally imprecise. For the 368 sampled counties for which statistics could be generated, the median value of the coefficient of variation of the percentage lacking *BPLS* was 51 percent.

The county-level estimates for the sampled counties were used in the subsequent regression analysis to compute model-based, indirect estimates for all counties in the United States. The county estimates were then aggregated to produce state-level estimates. The Hierarchical Bayes (HB) methodology used for these computations is described in chapter 4. The estimation method incorporates the standard error of the direct estimates into the model, thereby accounting for the imprecision of the estimates of the percentage lacking *BPLS*.

6.3 Indirect Estimates

The computation of indirect estimates for NALS followed the same approach as described in chapters 3 through 5 for the 2003 NAAL. The process involved the following:

- smoothing the direct relative variances (relvariances), as described in section 6.3.1;
- selecting predictor variables, as described in section 6.3.2;
- applying the HB model to produce the county estimates, as described in section 6.3.3; and
- evaluating the HB model, as described in section 6.3.4.

6.3.1 Smoothing the Direct Relative Variance Estimates

An assumption made with the HB model used is that the relative variances, or relvariances, of the county direct estimates of the percentages lacking *BPLS* were known. However, in practice only imprecise estimates of the relvariances were available. It was therefore necessary to smooth the direct relvariances prior to implementing the HB model, as described in section 4.2 and appendix A. The smoothing process used a two-step robust regression approach: the first step developed model-based estimates of proportions of adults lacking *BPLS*, that is, of the denominators of the relvariances; the second step developed model-based estimates of the relvariances of the county direct estimates of these percentages, incorporating the first-step model-based estimates of the proportions lacking *BPLS* in the second-step model.

In the first step, the logit of the proportion lacking *BPLS* in each sampled county was used as the dependent variable. Predictor variables were selected using a stepwise selection method. A robust regression M-estimation approach using the Stata (StataCorp, 2005) statistics package was used to arrive at the predicted values of the proportion lacking *BPLS* in each sampled county. Each county was assigned a weight equal to 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 square root was applied as an ad hoc method of approximating weighting by residual variance. Outliers (above 2 percent) were identified based on the bisquare function. The bisquare function is described in detail in the discussion on the RobustReg procedure in the SAS/STAT user’s manual (SAS Institute 2003). All the outliers were downweighted and none were set to zero.

The final model has the form:

$$log\,it\left(p_{ij}\right)=\gamma_0+\gamma_1 Z_{ij1}+\gamma_2 Z_{ij2}+\gamma_3 Z_{ij3}+\gamma_4 Z_{ij4}+\gamma_5 Z_{ij5}+\varepsilon_{ij} \tag{1}$$

where, in sampled county *j* in state *i,*

p_{ij}	=	the proportion of adults lacking *BPLS*;
Z_{ij1}	=	percentage of persons who are Black;
Z_{ij2}	=	percentage of persons for whom English is not a native language;
Z_{ij3}	=	percentage of persons in rural areas;
Z_{ij4}	=	percentage of persons who are Hispanic;
Z_{ij5}	=	percentage of persons age 25+ with a high school education or less; and
e_{ij}	=	the error term.

Table 6-1 presents the estimated parameters of the generalized variance function (GVF) model, processed on the 368 sampled counties, which had an R^2 value of .3; this is similar to the fit of the corresponding model for the 2003 data where the R^2 value was .4. The final model retained some variables that were not statistically significant because they were believed to be contributing to a better fit. As a check, the predicted values obtained from the model above were compared to the predicted values from the final HB model (after the final indirect estimates were produced). The correlation coefficient between the two sets of predicted values was .9; this value for 2003 data was also .9.

Table 6-1. Parameter estimates for the first step of the variance smoothing process for the county-level direct estimates of the percentage lacking Basic prose literacy skills: 1992

Parameter	Estimate	Standard error	95 percent confidence limits		Chi-square	p-value
γ_0	#	0.01	-0.06	#	5	.03
γ_1	0.1	0.04	0.03	0.18	8	.006
γ_2	0.1	0.08	#	0.30	4	.058
γ_3	#	0.02	-0.04	0.04	#	.924
γ_4	#	0.08	-0.17	0.14	#	.833
γ_5	0.5	0.07	0.40	0.67	59	<.001

Rounds to zero

SOURCE: U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

In the second step, the predicted value of the proportion lacking *BPLS* in each sample county from the regression model in equation (1) was used as a predictor variable in a generalized variance function (GVF) model to smooth the relvariance estimates. To make the relvariance model approximately linear in the parameters, a robust weighted least squares log-log model was used, where the weight was the square root of the degrees of freedom for the direct variance estimate. The robust regression approach was the same as that used in Step 1. The rationale for the use of the square root of degrees of freedom as weights in the GVF model was the same as that for the use of the square root of sample size in the regression model for predicting the proportion lacking *BPLS* (that is, the direct estimates of relvariances were weighted in the GVF regression by a measure that was related to their precision). This *ad hoc* weighting scheme downplayed the less precise relvariance estimates. The model has the form:

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

where, in sample county *j* in state *i*,

φ_{ij}^2 is the relvariance of the proportion of adults lacking *BPLS*;
$\tilde{p}_{ij}$ is the predicted proportion from model (1);
n_{ij} is the sample size; and
e_{ij} is the error term.

The outliers (5 percent) were identified based on the bisquare robust regression procedure. All the outliers were downweighted and none were set to zero. Table 6-2 contains the parameter estimates for the robust GVF regression processed on the 368 sampled counties. The model had an R^2 value of .4, the R^2 value obtained in the corresponding analysis of the 2003 data was also .4.

Table 6-2. Parameter estimates for the second step of the variance smoothing process for county-level relvariances: 1992

Parameter	Estimate	Standard error	95 percent confidence limits		Chi-square	p-value
η_0	#	1.76	-3.34	3.48	#	.993
η_1	-1.3	0.57	-2.44	-0.21	5	.023
η_2	-0.1	3.74	-7.25	7.44	#	.98
η_3	-1.1	0.09	-1.23	-0.90	156	<.001

Rounds to zero

SOURCE: U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

The predicted values of the relvariances for the sample percentages of adults lacking *BPLS* were computed based on the GVF regression model in equation (2) for all sample counties. These predicted values were then treated as known relvariances in the HB model.

6.3.2 Predictor Variables

The HB model for county and state predictions used county-level statistics, both as dependent and as predictor variables. To be considered for use in the HB model, county-level variables had to be available for all counties, whether or not they were included in the NALS sample, and also had to be consistently measured across counties.

All variables used in the final HB model for 2003 were considered in fitting the 1992 model, using definitions across the two years of the Census (1990 and 2000). The composite variable "percentage of the population that was Black or Hispanic", which resulted in the 2003 analysis, was divided back into its components (the percentage of the population that was Black, the percentage of the population that was Hispanic) for NALS. This allowed the stepwise model approach described below to provide a better fit to the county-level data by allowing the coefficients to differ between the two groups. Similarly, separate indicators of Census divisions used with the 2003 model were used in the 1992 model. Table 6-3 displays the variables that were considered for inclusion in the 1992 model, and table 6-4 presents correlations among the variables included in the model, which were computed on the 368 sampled counties.

A stepwise regression model, using a .05 level of significance, was used to select variables that had the greatest ability to explain the between-county variation in the percentage of adults lacking *BPLS*, using variables from the 1990 Census that matched those used in the 2003 model. Alternative specifications had little impact on overall model fit, and a final model was chosen based on parsimony. The income, foreign-born, and poverty variables were not significant when added to the model containing the other variables. The right-most column in table 6-3 displays the final predictor variables retained in the HB model. Two indicator variables were created for census divisions. The first indicator variable was equal to 1 if the county was in the New England division (and zero otherwise), and the second variable was equal to 1 if the county was in one of the North Central divisions (and zero otherwise).

Table 6-3. List of predictor variables, their source, and the selected predictor variables for the final model: 1992

Predictor variables	Source	Predictor variables used in final HB model
Percentage of the population who are foreign-born	1990 Census of Population	
Percentage of the population for whom English is not a native language	1990 Census of Population	X
Median household income	1990 Census of Population	
Percentage of the population age 25 and older with a high school education or less	1990 Census of Population	X
Percentage of the population who were Black	1990 Census of Population	X
Percentage of the population who were Hispanic	1990 Census of Population	X
Percentage of the population below the 150 percent poverty line	1990 Census of Population	
Indicator variables identifying census divisions[1]	1990 Census of Population	X
Indicator variable for counties in a State Adult Literacy Survey state	1992 National Adult Literacy Survey	X

[1] Two indicator variables related to Census divisions were included in the final model. The first indicator variable identified counties in the New England division, and the second variable identified counties in the North Central census divisions (combining the East and West North Central census divisions).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

Table 6-4. Correlation coefficients among predictor variables for the final small area model: 1992

Variable	Percentage Black	Percentage Hispanic	Percentage non-English speaking	New England census division	North Central census divisions	State Adult Literacy Survey sample indicator
Percentage with high school education or less	0.33	0.13	0.09	-0.09	-0.17	-0.06
Percentage Black		0.02	0.09	-0.07	-0.30	-0.18
Percentage Hispanic			0.85	-0.02	-0.29	0.13
Percentage non-English speaking				0.09	-0.35	0.14
New England census division					-0.13	-0.23
North Central census divisions						0.15

SOURCE: U.S. Department of Commerce, Census Bureau, Census 1990 Summary File 3. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

6.3.3 Model Development and Prediction for Counties and States

The Markov Chain Monte Carlo (MCMC) approach used with the 1992 data replicated the 2003 NAAL estimation approach described in section 4.3. The model was processed three times to obtain a total of 27,000 estimates of model parameter values that could be used to compute summary statistics for the resulting distribution of values, including their mean and credible intervals. The initial parameter estimates for the first run of the model were obtained using the ordinary least squares estimates for coefficients of the right-hand-side variables in the model. The final variance estimates of state and county variance terms that were obtained from the analysis of 2003 data were also used as starting values for the 1992 model, under the assumption these variances would be of similar magnitude across the two time periods, though ultimately these variances were estimated from the 1992 data. The second and third runs of the model shifted these starting values by 10 percent in each direction as a means of incorporating alternative assumptions regarding these variances. Trace plots were reviewed to check for independence and coverage, following the checks conducted for the 2003 model. The results for the HB regression model using the predictor variables displayed in table 6-1 are presented in table 6-5. The table shows the mean and median values of each estimated parameter across the MCMC draws, the standard deviation of these parameter estimates, and the upper and lower bounds on these estimates (corresponding to the 95 percent credible interval).[52]

[52] A scale reduction factor estimate ($\hat{R}$) was computed for the final set of 2003 estimates to assess the convergence of coefficient estimates across MCMC runs. The 1992 analysis was primarily replication of the model developed with the 2003 data and took the model specification and length of the MCMC runs as given to produce the estimates presented in this table without additional computation of the scale reduction factor estimate. In other words, the scale reduction is a measure of goodness of fit that is calculated after a model is estimated and is not an input into a model run, i.e., the scale reduction estimate from 2003 could not be reused in the 1992 model. In calculating estimates with the 1992 data, however, estimates of model variances from the 2003 model were used as starting values from which the 1992 model could begin its initial calculations.

Table 6-5. Regression coefficients and variances of random effects for the final HB model: 1992

Parameter	HB mean	HB Standard deviation	Median	95 percent credible interval Lower bound	Upper bound
Intercept	-3.3	0.17	-3.3	-3.64	-2.96
Percentage with high school education or less	4.6	0.53	4.6	3.46	5.65
Percentage Black	1.2	0.31	1.1	0.54	1.77
Percentage Hispanic	-0.4	0.73	-0.3	-1.83	1.00
Percentage non-English speaking	1.5	0.65	1.5	0.33	2.87
New England census division	0.1	0.24	0.1	-0.37	0.59
North Central census divisions	-0.1	0.09	-0.1	-0.31	0.08
State Adult Literacy Survey sample indicator	-0.1	0.09	-0.1	-0.31	0.04
Variance of county random effect	0.2	0.03	0.2	0.09	0.22
Variance of state random effect	#	0.01	#	#	0.04

Rounds to zero

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

The variables included in the model were selected based on their ability to predict percentage of adults lacking *BPLS* using a stepwise regression approach. The variables that are statistically significant in the HB model are the percentage of individuals in a county with a high school education or less, the percentage of individuals who were Black, and the percentage of individuals for whom English is not a native language. Other coefficients including the census division indicators and the SALS sample indicator, are not statistically significant. The county and state random effect variables are significant with the county effect dominating the state effect, indicating higher significant variations in percentage of adults lacking *BPLS* among counties than among states.

The parameter estimates from each of the MCMC samples were used to create posterior distributions of indirect estimates of the percentages of adults lacking *BPLS* for individual counties, whether or not they were included in the NALS sample. The approach to creating these estimates described in section 4.4 was applied with the 1992 data using the approach described in 4.4.3.[53] The county-level estimates are provided at the NAAL website (http://nces.ed.gov/NAAL). The state estimates are presented in appendix C and figure C-1 provides a graph of the estimates.

[53] Resource constraints limited the ability to develop an interactive, web-based tool to compare differences in estimates from the 1992 data developed for pairs of states or pairs of counties within states. A user can, however, make these comparisons using the formula presented in section 4.6 to estimate the standard error of the difference in point estimates between any two states or pair of counties.

Given the final model and its predictions, it is possible to characterize the precision of the estimates in terms of the width of credible intervals for the predicted percentages of adults lacking *BPLS* and also the coefficient of variation (CV) of these estimates. Table 6-6 provides a summary of the variability of the 1992 HB estimates.

Table 6-6. Distribution of credible interval widths and coefficients of variation for county and state estimates: 1992

Statistic	Percentile				Median
	20	40	60	80	
County estimates					
95 percent credible interval width (percent)	13.3	16.4	21.0	28.7	18.2
Coefficient of variation (percent)	30.0	33.5	35.7	37.6	34.7
Sampled county estimates					
95 percent credible interval width (percent)	9.5	11.2	13.3	16.4	12.2
Coefficient of variation (percent)	22.5	27.2	31.4	36.1	29.2
Nonsampled county estimates					
95 percent credible interval width (percent)	14.1	17.3	22.4	29.6	19.3
Coefficient of variation (percent)	31.0	34.0	36.0	37.7	35.0
State estimates					
95 percent credible interval width (percent)	5.0	6.2	7.0	7.7	6.5
Coefficient of variation (percent)	12.1	14.5	15.8	17.9	15.3

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.

As expected, the table shows that county estimates are more precise for counties with NALS sample cases—median credible interval width of 12 percent—than for counties not in the NALS sample—median width of 19 percent.

The coefficients of variation (CV) in table 6-6 are obtained by dividing the standard deviation of the estimates for a given county across all draws of the Monte Carlo simulation by the mean of the MCMC county estimates. Most of the CVs for the indirect county estimates are on the order of 30 percent or more. Half of the counties have a CV of more than 35 percent. An estimate of the percentage of adults lacking *BPLS* with a CV of this magnitude is highly imprecise. It is important for the users of these county estimates to recognize this fact and treat the estimates with due caution. The state estimates are more precise, in that their CVs are smaller than county-level estimates, but the median CV is still relatively large at 15 percent.

6.3.4 Model Evaluation

There are two approaches to evaluating the 1992 results. The first (presented in this section) evaluates the ability of the model to fit the 1992 data and therefore is internal to the 1992 data.[54] The second is comparison of the goodness of fit of the model in 1992 to the model developed in 2003. This second comparison is presented in chapter 7.

Three measures were computed to assess the goodness of fit of the 1992 HB model, all based on comparisons of the direct to indirect MCMC estimates (see section 5.1 for the application of these measures to the 2003 data). The measures are the following:

- A 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 indirect county estimates. The posterior predictive *p* value is the proportion of the samples that had a smaller simulated discrepancy measure (as opposed to the direct discrepancy measure) and should be close to 0.5 if the model fits the data well. Since the distribution of the estimates of the percentage lacking *BPLS* differs from a normal distribution and since the estimates were less than 10 percent for several counties, 9 percent of the simulated estimates were negative and were therefore excluded. After these exclusions, the *p* value for the final model was equal to .65.

- A county-level measure computed as the proportion of the 27,000 MCMC samples that had a smaller simulated value than the direct estimates for the county. These proportions are expected to vary across the counties, but to even out across counties so that there should be a small number of counties with values close to 0 or 1. Across the counties, this proportion ranged from .07 to .94, with a global average of .5. There were 4 extreme values for counties, with proportions of .07, .10, .87, and .94.

- A 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 MCMC samples. Values of this measure are expected to mostly range between, say, -1.96 and 1.96 with an overall average of around 0. The values of this statistic for the 1992 HB county estimates ranged from -2.21 to 1.89, with a global average of -0.04.

---

[54] The main objective of this task was to produce model-based estimates for the 2003 NAAL and to replicate the same methodology for the 1992 NALS to arrive at comparable models. Therefore, the 1992 NALS modeling was not carried out at the same intensity as the 2003 models.

The above measures indicate that the model did not tend to systematically overestimate or underestimate the value of the percentages of adults lacking *BPLS* relative to the direct estimates in 1992.

Also, an evaluation of the relationship between the indirect and direct estimates was conducted. The correlation (*r* =0.98) was quite high for those counties that had 100 or more sampled individuals and fell to .6 for counties with fewer than 20 sampled individuals as expected.

7. Comparison of the 1992 and 2003 Indirect County and State Estimates[55]

Chapters 2 through 6 apply Hierarchical Bayes (HB) regression modeling to two national assessments of adult literacy to produce county and state indirect estimates of the percentage lacking *BPLS* in 1992 and 2003. A model was initially developed and tested with data from the 2003 National Assessment of Adult Literacy (NAAL). Once a final HB model was developed for the NAAL, the same estimation method was applied to the 1992 National Adult Literacy Survey (NALS). The estimation approach for the 1992 NALS was the same as that used for the 2003 NAAL for the following reasons.

- The 2003 NAAL had more predictor variables available. An extensive search for key predictors was conducted, which resulted in retaining variables from the 2000 Census only.
- There are not as many predictor variables available for 1992. Given the findings from the 2003 modeling, the search for predictor variables were focused on data available from Census 1990 that were related to those variables used in 2003 modeling.
- For the sake of comparability between years, there is much value in using the same model (i.e., same model structure, and same or similar predictor variables).

As discussed below, the estimates from the two years are generally comparable in their precision, though as expected, by applying a model developed to provide the best fit to 2003 data, the credible intervals are wider for 1992.

The objective of this chapter is to compare the indirect county and state estimates between 1992 and 2003. First, however, by way of background, section 7.1 compares the HB models for the two years and examines their fit to the data. Section 7.2 then provides guidelines on how to compare indirect estimates for each county or state for 1992 with the corresponding 2003 indirect estimate using a web tool available on the NCES website.

---

[55] Following authors contributed to this chapter: Dan Sherman and Jennifer Dillman, American Institutes for Research, and Tom Krenzke, Westat.

7.1 Comparison of the 1992 and 2003 HB Models

A summary of the comparisons made between the 1992 and the 2003 HB models are given below.

- A comparison of the 1992 estimated model parameters (table 6-5) with the associated estimates for the 2003 NAAL data (table 4-2) indicates that there are predictor variables available from the Population Census that predict the county percentages lacking *BPLS* in both years. The estimated variances of county and state random effects are similar for the two years (the HB mean for the county random effect was 0.15 in the 1992 model and 0.12 in the 2003 model; the mean of the state random effect was 0.01 for 1992 and 0.02 for 2003). The larger county random effects indicate that there is likely to be larger variation among counties than among states. Both models include variables relating to education attainment, race/ethnicity, indicators for census divisions, and state assessment indicators. Foreign-born status was used in the 2003 model only, while native English speaking status was used in the 1992 model only.

- The simple correlation between the direct county-level estimates of the percentage lacking *BPLS* and their predicted values for both years is highest for counties with the larger sample sizes. For smaller counties, there is a great deal of variability in predictions with decreasing sample size, as expected.

- The county indirect estimates of the percentage lacking *BPLS* are subject to substantial variability for both 1992 (table 6-6) and 2003 (table 4-3). The median of the 95 percent credible interval width for all counties is 18 percent in 1992 and 15 percent in 2003. The state estimates were more precise: the median widths of the 95 percent credible intervals are 7 percent in 1992 and 6 percent in 2003.

- The 95 percent credible interval widths are wider for counties not in the sample than for counties included in the sample. The median interval width for sample counties in 1992 was 12 percent and also 12 percent in 2003; for counties not in the sample, the median interval was 19 percent in 1992 and 15 percent in 2003.

- The coefficients of variation (CV) for estimates of the percentage lacking *BPLS* are similar across the two assessments. For sampled counties, the CV is 29 percent in 1992 and 28 in 2003. For other counties, the CV of estimates is 35 percent in 1992 and 33 percent in 2003. For state estimates, the median CV is 15 percent in 1992 and 14 percent in 2003.

These findings indicate that, overall, the county estimates for both years have coefficients of variation on the order of 30 percent or more.[56] Thus, for example, an approximate

[56] The finding that 1992 estimates were less precise may reflect the fact the analysis of 2003 data involved extensive consideration of alternative explanatory variables and model specifications (chapter 5). The analysis of 1992 data primarily sought to apply the HB modeling to the

95 percent credible interval for a county with an estimated 14 percent of adults lacking *BPLS* (i.e., approximately the national average for both years) and a CV of 35 percent is from 4 percent to 24 percent.[57]

The state estimates are more precise, with median coefficients of variation of 15 percent in 1992 and 14 percent in 2003. However, with a CV of this magnitude, the estimate is still relatively imprecise. For example, for a state with an estimated 14 percent lacking *BPLS* with a CV of 15 percent, the 95 percent credible interval is from 10 to 18 percent.

7.2 Comparisons of the 1992 and 2003 Indirect Estimates

Since the HB models were fit using both the 1992 and 2003 assessment data, comparisons can be made between the indirect estimates for any given county or state for the two years (excluding three 1992 counties that did not exist in 2003 and three 2003 counties that did not exist in 1992). Across all counties, the correlation between the county estimates for the two years was .8. For states, the correlation was .7.

The approximate method of creating credible intervals described in section 4.6 has been used to create credible intervals of the differences between the 1992 and 2003 county and state indirect estimates. These credible intervals are available at the NAAL website via a web tool similar to the one created for the 2003 estimates, as described in section 4.6.

When comparing the 1992 and 2003 estimates for a single specific county or state, a credible interval for the difference that does not include zero indicates that the two estimates are different with the probability of .95. For any specific comparison, there is a 5 percent statistical risk of obtaining a credible

---

1992 NAAL data. The 2003 estimates may be considered optimal in the sense that are based on a model that was fitted using alternative sets of variables, with a final model chosen based on goodness-of-fit considerations. It is possible that a comparable effort with 1992 data might have produced more precise estimates, though the analysis of 2003 data (table 5-2) suggests that model estimates across alternative specifications using available variables were highly correlated.

[57] This approximate interval is computed as follows. Since the CV is equal to the standard error divided by the point estimate, then the standard error for this example is equal to .35 * .14 = .049. The lower bound of a 95 percent confidence interval is computed as .14-1.96 * .049, which is equal to .044. The upper bound is computed as .14 + 1.96 * .049, which is equal to .236. Therefore, an approximate 95 percent credible interval is from 4 percent to 24 percent.

interval that does not include zero when there is in fact no true difference between the population percentages for 1992 and 2003 (i.e., a Type I error). When multiple comparisons are made, the risk of making a Type I error with one or more of the comparisons can be substantial. To avoid this situation, users need to carefully select the specific comparisons they make when utilizing the tool. To focus the user on specific comparisons, the tool is constructed to allow only one comparison at a time.

As noted in section 7.1, the indirect estimates of the percentage lacking *BPLS* have low precision for each year. This imprecision of county estimates results in the underlying estimates of change between survey years for individual counties also being imprecise. This could be related to the finding that few county comparisons between 1992 and 2003 are significant, i.e., show detectable differences between 1992 and 2003 estimates (at the .05 level). Using the approximate method described in section 4.6, 1 percent of all counties exhibit significant differences between their indirect estimates across the two years.

The small area estimation approach was used to create indirect estimates because there is no data source available that can provide reliable direct estimates of the percentage of adults at the lowest literacy level for all counties and states in the nation. Users need to be aware of the credible intervals associated with the indirect estimates, as they gain a general picture of the literacy status for all counties and states.

As mentioned above, the state estimates are more precise than the county estimates, enabling better detection of statistically significant change for individual states. Using the approximate method, 9 percent of all states exhibit significant differences between their indirect estimates for 1992 and 2003.

REFERENCES

American Institutes for Research and Cohen, J. (2006). AM Statistical Software Beta Version 0.06.03. Washington, D.C.

Bayarri, M.J. and Berger, J. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*: 1127-1142.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 5*: 279-292.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F.M. Lord, M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing.

Bock, R. D., and Aitkin, M. (1982). Marginal Maximum Likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*: 443-459.

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

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

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

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. 2nd edition. Texts in statistical science. Chapman and Hall/CRC.

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

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.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST.* An Official Journal of the Spanish Society of Statistics and Operations Research. Vol. 15, No. 1 pp. 1-96.

Kirsch, I., Yamamoto, K., Norris, N., Rock, D. Jungeblut, A. O'Reilly, P., Berlin, M., Mohadjer, L. Waksberg, J.L., Goksel, H., Burke, J., Rieger, S., Green, J., Klein, M., Campbell, A., Jenkins, L.

Kolstad, A., Mosenthal, P., and Baldi, S. (2000). *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (NCES 2001-457). 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.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Eribaum Associates.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS—A Bayesian modeling framework: Concepts, structure, and extensibility, Statistics and Computing, *10*(4): 325-337.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*: 133-161.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*: 131-154.

Mislevy, R. J. and Sheehan, K. M. (1987). Marginal estimation procedures. In A. Beaton (Ed.), *The NAEP 1983-84 Technical Report* (pp. 293 - 360). Princeton, NJ: Education Testing Service.

Rao, J.N.K. (2003). *Small area estimation*. Wiley-Interscience.

Robert, C.P., and Casella, G. (1999). Monte Carlo Statistical Methods. New York: Springer-Verlag.

SAS Institute (2003). SAS version 9.1. Cary, NC: SAS Institute, Inc. SAS OnlineDoc 9.1.3: http://support.sas.com/onlinedoc/913/docMainpage.jsp.

Spiegelhalter D. J., N. 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.

StataCorp. (2005). Stata Statistical Software: Release 9. College Station, TX: StatCorp LP.

White, S. and Dillow, S. (2005). *Key Concepts and Features of the 2003 National Assessment of Adult Literacy* (NCES 2006-471). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

You, Y. and Rao, J.N.K. (2002) Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics, 30*: 3 – 15.

Appendix A

This page left intentionally blank.

2003 NAAL PREDICTOR VARIABLE SOURCES[58]

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. This appendix contains a list of data sources used to extract county-level and state-level data for the 2003 NAAL small area models. A listing of all county variables considered for modeling is given in table A-1. The listing is sorted by major variable type (i.e., poverty, income, etc). The county level predictor variable sources considered for 2003 NAAL were as follows.

Census 2000 Data[59]: Summary File 3 (SF3) was used to extract county level variables. The SF3 contains the "short form" items (items asked of all households) and includes information about age, sex, race, Hispanic or Latino origin, household relationship, and owner/renter status. The SF3 also contains the "long form" data coming from questions asked of about one-sixth of America's households. The questions include such topics as income, education, language spoken, housing structure, housing costs, commuting.

Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program[60]: The Census Bureau, with support from other Federal agencies, created the SAIPE program to provide more current small area estimates of selected income and poverty statistics than the most recent decennial census.

Bureau of Labor Statistics (BLS)[61]: The Local Area Unemployment Statistics (LAUS) program produces monthly and annual employment, unemployment, and labor force data for Census regions and divisions, states, counties, metropolitan areas, and some cities, by place of residence.

Bureau of Economic Analysis (BEA)[62]: The BEA prepares estimates of personal income for local areas (counties, metropolitan areas, and the BEA economic areas). The personal income of an area is the income that is received by, or on behalf of, the residents of that area.

U.S. Department of Agriculture (USDA)[63]: The USDA Economic Research Service provides codes that classify each county according to metro and non-metro classifications. Here is the description from the

[58] This appendix was written by Tom Krenzke and Lin Li, Westat.

[59] The website for the Census 2000 SF3 is http://www.census.gov/Press-Release/www/2002/sumfile3.html.

[60] The website for the SAIPE program is http://www.census.gov/hhes/www/saipe/.

[61] The website for the LAUS program is http://www.bls.gov/lau/.

[62] The website for the Bureau of Economic Analysis is http://www.bea.gov/.

USDA website. "The 2003 Rural-urban Continuum Codes form a classification scheme that distinguishes metropolitan counties by size and nonmetropolitan counties by degree of urbanization and proximity to metro areas. The standard Office of Management and Budget (OMB) metro and nonmetro categories have been subdivided into three metro and six nonmetro categories, resulting in a 9-part county codification."

A listing of the state-level variables that were considered for the small area models is contained in table A-2. After accumulating the list, state-level variables were not given further consideration if it had a closely associated subject matter and form to a county-level variable candidate. All other literacy-related variables were downloaded or key entered, and these variables are flagged in table A-2. Once obtained, the variables entered a stepwise regression selection process and proceeded through the variable selection process as described in chapter 3. A description of the sources for the state-level covariates follows.

American Community Survey (ACS)[64]: The ACS is a nationwide survey that is designed to provide data on communities in years between the decennial censuses. The ACS replaces the Census long form and is a critical element in the Census Bureau's reengineered 2010 census plan.

Other Census Bureau programs[65]: Besides the ACS, other data from the Census Bureau was collected from the following programs: Population Estimates, Public Employment and Payroll Data, Current Population Survey, Current Population Reports, Federal Aid to States for Fiscal Year, State and Local Government Finance Estimates, and data on housing vacancies and home ownership from the Housing Vacancy Survey.

Bureau of Labor Statistics[66]: Besides the BLS LAUS program mentioned above, state-level data were considered from the Current Employment Statistics Program, which surveys over 160,000 businesses and government agencies each month. The Employment and Wages annual averages were also included in the selection process.

---

[63] The website for the U.S. Department of Agriculture is http://www.usda.gov/wps/portal/usdahome.

[64] The website for the American Community Survey is http://www.census.gov/acs/www/.

[65] The following are the websites for the other Census Bureau programs: for the Census Population Estimates: http://www.census.gov/popest/estimates.php; Public Employment and Payroll Data: http://www.census.gov/govs/www/apes.html; Current Population Survey: http://www.census.gov/cps/; Current Population Reports: http://www.census.gov/main/www/cprs.html; Federal Aid to States for Fiscal Year data: http://www.census.gov/prod/2004pubs/03fas.pdf; State and Local Government Finance Estimates: http://www.census.gov/govs/www/financegen.html; Housing Vacancy Survey: http://www.census.gov/hhes/www/housing/hvs/hvs.html.

[66] The website for Bureau of Labor Statistics is http://www.bls.gov/.

Behavioral Risk Surveillance System (BRFSS)[67]. This program tracks health risks in the United States and was established by Centers for Disease Control and Prevention (CDC). Information (e.g., tobacco use, disability, exercise) from the survey is used to improve the health of the American people.

Adult Education Data[68]: The Office of Vocation and Adult Education (OVAE) collects data on adult education program enrollments from each state. Unpublished sampling frame data for the 2003 Adult Education Program Survey was considered for the small area models.

The Integrated Postsecondary Education Data System (IPEDS)[69]: This NCES program collects data through a system of surveys from primary providers of postsecondary education.

Other sources[70]: State-level data from other sources were obtained. This was done primarily through the use of the Statistical Abstract of the United States, which is a guide to sources of other data from the Census Bureau, other Federal agencies, and private organizations. These other sources included National Highway Safety Traffic Administration's *Traffic Safety Facts,* the BEA's *Survey of Current Business,* National Center for Health Statistics' *Vital Statistics of the United States,* the American Medical Association's *Physician Characteristics and Distribution in the U.S.,* National Education Association's *Estimates of School Statistics Database*, the Federal Bureau of Investigation's *Crime in the United States,* and the Energy Information Administration's *State Energy Data Report.*

---

[67] The website for the BRFSS program is http://www.cdc.gov/brfss/.

[68] The website for OVAE is http://www.ed.gov/about/offices/list/ovae/pi/AdultEd/index.html.

[69] The website for the IPEDS program is http://nces.ed.gov/ipeds/.

[70] The following are the websites for the other sources: *Traffic Safety Facts:* http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSFAnn/2003HTMLTSF/TSF2003.htm; *Survey of Current Business:* http://www.bea.gov/scb/index.htm; *Vital Statistics of the United States:* http://www.cdc.gov/nchs/products/pubs/pubd/vsus/vsus.htm; American Medical Association: http://www.ama-assn.org/; *Estimates of School Statistics Database:* http://www.nea.org/edstats/RankFull06b.htm; *Crime in the United States:* http://www.fbi.gov/ucr/cius_03/pdf/toc03.pdf; Energy Information Administration: http://www.eia.doe.gov.

Table A-1. Listing of county-level variables considered in the variable selection process: 2003

County characteristics	Source	Year
Poverty		
Percent below 150 percent poverty line	SF3	2000
Percent in poverty	SF3	1999
All ages in poverty	SAIPE	2003
Income		
Median household income	SF3	1999
Median household income	SAIPE	2003
Per capita personal income	BEA	2003
Education		
Percent of people age 25+: with education less than high school	SF3	2000
Percent of people age 25+: with high school diploma, no college	SF3	2000
Percent of people age 25+: with high school diploma or less	SF3	2000
Percent of people age 25+: with more than high school	SF3	2000
English-speaking ability for people who speak other language		
Percent of people age 5+: speak other language and speak English not at all or not well	SF3	2000
Percent of people age 5+: speak other language and speak English well or very well	SF3	2000
Urban/rural		
Percent of people inside or outside urbanized area	SF3	2000
Percent of people in rural farm or nonfarm area	SF3	2000
Counties in metro area of 1 million population or more	USDA	2000
Counties in metro areas of less than 1 million population	USDA	2000
Non-metro counties	USDA	2000
Race/ethnicity		
Percent of Hispanics	SF3	2000
Percent of Blacks	SF3	2000
Percent of Asians	SF3	2000
Percent of Native Americans	SF3	2000
Percent of Other	SF3	2000
Length of stay for foreign-born people		
Percent of foreign-born people who stayed in U.S. for 5 years or less	SF3	2000
Percent of foreign-born people who stayed in U.S. for 6 to 20 years	SF3	2000
Percent of foreign-born people who stayed in U.S. for 20 years or less	SF3	2000
Percent of foreign-born people who stayed in U.S. for 21 years or more	SF3	2000
Age		
Percent of people 16-54 years old	SF3	2000
Percent of people 55-64 years old	SF3	2000
Percent of people 65+ years old	SF3	2000
Gender		
Percent male age 16+	SF3	2000

See notes at end of table.

Table A-1. Listing of county-level variables considered in the variable selection process: 2003—Continued

County characteristics	Source	Year
Employment status		
Unemployment rate	BLS	2003
Percent of people age 20-64: in armed forces	SF3	2000
Percent of people age 20-64: in labor force and employed	SF3	2000
Percent of people age 20-64: in labor force and unemployed	SF3	2000
Percent of people age 20-64: no in labor force	SF3	2000
Occupation		
Percent management/professional occupations	SF3	2000
Percent service occupation	SF3	2000
Percent sales/office occupation	SF3	2000
Percent farming/fishing/forestry occupation	SF3	2000
Percent construction/extraction/maintenance occupation	SF3	2000
Percent production/transportation/moving occupation	SF3	2000
Census division		
New England	SF3	2000
Middle Atlantic	SF3	2000
East North Central	SF3	2000
West North Central	SF3	2000
South Atlantic	SF3	2000
East South Central	SF3	2000
West South Central	SF3	2000
Mountain	SF3	2000
Pacific	SF3	2000
Journey to work:		
Percent of less than 30 minutes to work	SF3	2000
Percent of less than 30 minutes to work by public transportation	SF3	2000
Percent of 30-44 minutes to work	SF3	2000
Percent of 30-44 minutes to work by public transportation	SF3	2000
Percent of 45-59 minutes to work	SF3	2000
Percent of 45-59 minutes to work by public transportation	SF3	2000
Percent of 30-59 minutes to work	SF3	2000
Percent of 30-59 minutes to work by public transportation	SF3	2000
Percent of 60+ minutes to work	SF3	2000
Percent of 60+ minutes to work by public transportation	SF3	2000
Percent of people 16+ worked in state of residence	SF3	2000
Percent of people 16+ worked in county of residence	SF3	2000
Ancestry:		
Percent of Arab ancestry	SF3	2000
Percent of eastern European ancestry	SF3	2000
Percent of European ancestry	SF3	2000
Percent of northern European ancestry	SF3	2000
Percent of subsaharan African ancestry	SF3	2000
Percent of west Indian ancestry	SF3	2000

See notes at end of table.

Table A-1. Listing of county-level variables considered in the variable selection process: 2003—Continued

County characteristics	Source	Year
Housing unit tenure and phone service:		
Percent of owner occupied housing unit	SF3	2000
Percent of renter occupied housing unit	SF3	2000
Percent of owner occupied housing unit with phone service available	SF3	2000
Percent of renter occupied housing unit with phone service available	SF3	2000
Percent of occupied housing unit	SF3	2000
Plumbing facilities		
Percent of housing unit with plumbing facilities	SF3	2000
Marital status:		
Percent of people 15+ never married	SF3	2000
Percent of people 15+ married	SF3	2000
Percent of people 15+ widowed	SF3	2000
Percent of people 15+ divorced	SF3	2000
Migration:		
Percent of people 5+ in different house in 1995	SF3	2000
Percent of people 5+ in different house and in USA in 1995	SF3	2000
Percent of people 5+ in different county in 1995	SF3	2000
Percent of people 5+ in different state in 1995	SF3	2000
Employment disability		
Percent of 16-64 years old: with employment disability	SF3	2000
Design Variable		
State Assessment of Adult Literacy identifier	NAAL	2003

NOTE: The acronyms are SF3 = Summary File 3 from Census 2000; BEA = Bureau of Economic Analysis; BLS = Bureau of Labor Statistics; NAAL = National Assessment of Adult Literacy; SAIPE = Small Area Income and Poverty Estimates Program; USDA = U.S. Department of Agriculture.

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3; U.S. Department of Agriculture, Economic Research Service (2000); U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy. The website for the SAIPE program is http://www.census.gov/hhes/www/saipe/. The website for the BLS Local Area Unemployment Statistics program is http://www.bls.gov/lau/. The website for the Bureau of Economic Analysis is http://www.bea.gov/.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003

State characteristics	Source	Year	
Poverty			
Poverty percent all ages	SAIPE	2003	
Percent of Persons Below Poverty Level	ACS	2003	
Percent of Children Under 18 Years Below Poverty Level in the Past 12 Months (For Whom Poverty Status is Determined)	ACS	2003	
Percent of People 65 Years and Over Below Poverty Level in the Past 12 Months	ACS	2003	
Percent of People Below Poverty Level in the Past 12 Months (For Whom Poverty Status is Determined)	ACS	2003	
Percent of Related Children Under 18 Years Below Poverty Level in the Past 12 Months	ACS	2003	
Income			
Per capita income	BEA	2003	
Median household income	SAIPE	2003	
Median Earnings for Female Full-Time, Year-Round Workers (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Median Earnings for Male Full-Time, Year-Round Workers (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Median Family Income (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Median Household Income (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Percent of Households With Cash Public Assistance Income	ACS	2003	
Percent of Households With Retirement Income	ACS	2003	
Personal Income Per Capita in Constant (2000) Dollars	U.S. BEA, *Survey of Current Business,* April 2005.	2004	
Average Annual Pay	U.S. BLS, ‘‘Employment and Wages, Annual Averages,’’ 2002 and 2003.	2003	
Education			
Percent of People 25 Years and Over Who Have Completed High School (Including Equivalency)	ACS	2003	
Percent of People 25 Years and Over Who Have Completed a Bachelor's Degree	ACS	2003	
Percent of People 25 Years and Over Who Have Completed an Advanced Degree	ACS	2003	
Percent of People 25 Years and Over with Bachelor's Degree or More	CPS	2004	
Adult Basic Education Enrollment	OVAE	2001	*
Adult Secondary Education Enrollment	OVAE	2001	*
English as a Second Language Enrollment	OVAE	2001	*
Total Adult Education Enrollment	OVAE	2001	*
Graduation rate	IPEDS	2003	*
Public Elementary and Secondary School Teachers' Average Salaries	National Education Association, Washington, DC, Estimates of School Statistics Database.	2004	
Instructor salary	IPEDS	2003	*

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year	
Education (continued)			
Average financial aid	IPEDS	2003	*
Annual college cost	IPEDS	2003	*
English-speaking ability			
Percent of People 5 Years and Over Who Speak English Less Than "Very Well"	ACS	2003	
Percent of People 5 Years and Over Who Speak Spanish at Home	ACS	2003	
Percent of People 5 Years and Over Who Speak a Language Other Than English At Home	ACS	2003	
Race/ethnicity			
Percent of the Total Population Who are American Indian and Alaska Native Alone	ACS	2003	
Percent of the Total Population Who are Asian Alone	ACS	2003	
Percent of the Total Population Who are Black or African American Alone	ACS	2003	
Percent of the Total Population Who are Native Hawaiian and Other Pacific Islander Alone	ACS	2003	
Percent of the Total Population Who are Some Other Race Alone	ACS	2003	
Percent of the Total Population Who are Two or More Races	ACS	2003	
Percent of the Total Population Who are Two or More Races Excluding Some Other Race	ACS	2003	
Percent of the Total Population Who are White Alone	ACS	2003	
Percent of the Total Population Who are White Alone, Not Hispanic or Latino	ACS	2003	
White Population Alone, Number	U.S. Census Bureau, ‘‘Table 4: Annual Estimates of the Population by Race Alone and Hispanic or Latino Origin for the United States and States: July 1, 2004 (SC-EST2004-04)’’.	2004	
Black or African American Population Alone, Number	Same as above	2004	
American Indian, Alaska Native Population Alone, Number	Same as above	2004	
Asian Population Alone, Number	Same as above	2004	
Native Hawaiian and Other Pacific Islander Population Alone, Number	Same as above	2004	
Two or More Races Population, Number	Same as above	2004	
White Population Alone, Percent	Same as above	2004	
Black or African American Population Alone, Percent	Same as above	2004	
American Indian, Alaska Native Population Alone, Percent	Same as above	2004	
Asian Population Alone, Percent	Same as above	2004	
Native Hawaiian and Other Pacific Islander Population Alone, Percent	Same as above	2004	
Two or More Races Population, Percent	Same as above	2004	
Hispanic or Latino Origin Population, Number	Same as above	2004	
Hispanic or Latino Origin Population, Percent	Same as above	2004	
Non-Hispanic White Alone Population, Number	Same as above	2004	
Non-Hispanic White Alone Population, Percent	Same as above	2004	

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year	
Country of birth			
Percent of People Born in Asia	ACS	2003	
Percent of People Born in Europe	ACS	2003	
Percent of People Born in Latin America	ACS	2003	
Percent of People Born in Mexico	ACS	2003	
Percent of People Who are Foreign Born	ACS	2003	
Age			
Age Dependency Ratio of the Total Population	ACS	2003	
Median Age of the Total Population	ACS	2003	
Percent of the Total Population Who are 65 Years and Over	ACS	2003	
Percent of the Total Population Who are 85 Years and Over	ACS	2003	
Percent of Children Under 6 Years Old With All Parents in the Labor Force	ACS	2003	
Percent of Households That are Married-Couple Families With Own Children Under 18 Years	ACS	2003	
Percent of Households With One or More People Under 18 Years	ACS	2003	
Percent of Grandparents Living With Grandchildren and Responsible for Them	ACS	2003	*
Percent of Households With One or More People 65 Years and Over	ACS	2003	
Population Under 18 Years Old	U.S. Census Bureau, ‘‘Population estimates by State, Age and Sex for States and for Puerto Rico: April 1, 2000 to July 1, 2004’’	2004	
Population 65 Years Old and Over	Same as above	2004	
Gender			
Sex Ratio of the Total Population	ACS	2003	
Percent of employment	BLS	2003	
Employment			
Employment/Population Ratio for the Population 16 to 64 Years Old	ACS	2003	
Percent of Married-Couple Families With Both Husband and Wife in the Labor Force	ACS	2003	
Percent of People 16 Years and Over Who are in the Labor Force (Including Armed Forces)	ACS	2003	
State Government Full-Time Equivalent Employment Per 10,000 Population	U.S. Census Bureau; *Public Employment and Payroll Data.*	2003	

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year
Employment (Continued)		
Unemployment Rate	U.S. Bureau of Labor Statistics, Local Area Unemployment Statistics, *Geographic Profile of Employment and Unemployment, 2004 Annual Averages.*	2004
Nonfarm Employment—Percent in Manufacturing	U.S. Bureau of Labor Statistics, the *Current Employment Statistics* program.	2004
Percent of Civilian Employed People 16 Years and Over Who Were Private Wage and Salary Workers	ACS	2003
Occupation		
Percent of Civilian Employed People 16 Years and Over in Management, Business and Financial Operations Occupations	ACS	2003
Percent of Civilian Employed People 16 Years and Over in Professional and Related Occupations	ACS	2003
Percent of Civilian Employed People 16 Years and Over in Service Occupations	ACS	2003
Percent of Civilian Employed People 16 Years and Over in the Information Industry	ACS	2003
Percent of Civilian Employed People 16 Years and Over in the Manufacturing Industry	ACS	2003
Journey to work		
Mean Travel Time to Work of Workers 16 Years and Over Who Did Not Work at Home (Minutes)	ACS	2003
Percent of Workers 16 Years and Over Who Traveled to Work by Car, Truck, or Van—Carpooled	ACS	2003
Percent of Workers 16 Years and Over Who Traveled to Work by Car, Truck, or Van—Drove Alone	ACS	2003
Percent of Workers 16 Years and Over Who Traveled to Work by Public Transportation (Including Taxicab)	ACS	2003
Percent of Workers 16 Years and Over Who Worked Outside County of Residence	ACS	2003
Housing unit tenure		
Percent of Occupied Housing Units That are Owner-occupied	ACS	2003
Homeownership Rate	*Housing Vacancies and HomeOwnership* (CPS/HVS)	2004

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year	
Housing Financial Characteristics			
Median Housing Value of Specified Owner-occupied Housing Units (In 2003 Inflation-adjusted Dollars)	ACS	2003	*
Median Monthly Housing Costs for Specified Owner-occupied Housing Units With a Mortgage (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Median Monthly Housing Costs for Specified Renter-occupied Housing Units (In 2003 Inflation-adjusted Dollars)	ACS	2003	
Percent of Mortgaged Owners Spending 30 Percent or More of Household Income on Selected Monthly Owner Costs	ACS	2003	
Percent of Specified Renter-occupied Units Spending 30 Percent or More of Household Income on Rent and Utilities	ACS	2003	
Other Housing Characteristics			
Percent of Housing Units That Were One-Unit Detached	ACS	2003	
Percent of Housing Units That Were Mobile Homes	ACS	2003	
Percent of Housing Units That Were Built in 1939 or Earlier	ACS	2003	
Percent of Housing Units That Were Built in 2000 or Later	ACS	2003	
Percent of Housing Units That are Mobile Homes	ACS	2003	
Percent of Occupied Housing Units With Electricity as Principal Heating Fuel	ACS	2003	
Percent of Occupied Housing Units With Fuel Oil, Kerosene, Etc. as Principal Heating Fuel	ACS	2003	
Percent of Occupied Housing Units With Gas as Principal Heating Fuel	ACS	2003	
Percent of Occupied Housing Units That Were Moved into in 2000 or Later	ACS	2003	
Percent of Occupied Housing Units With 1.01 or More Occupants Per Room	ACS	2003	
Average Household Size	ACS	2003	*
Marital status			
Percent of Men 15 Years and Over Who Were Never Married	ACS	2003	
Percent of Women 15 Years and Over Who Were Never Married	ACS	2003	
Percent of Households That are Married-Couple Families	ACS	2003	
Migration			
Percent of People 1 Year and Over Who Lived in a Different House Within the Same State 1 Year Ago	ACS	2003	
Percent of People 1 Year and Over Who Lived in a Different House in the U.S. 1 Year Ago	ACS	2003	
Percent of People 1 Year and Over Who Lived in a Different State 1 Year Ago	ACS	2003	
Percent of the Native Population Born in their State of Residence	ACS	2003	
Disability			
Percent of People 21 to 64 Years Old With a Disability	ACS	2003	
Percent of People 5 to 20 Years Old With a Disability	ACS	2003	
Percent of People 65 Years and Over With a Disability	ACS	2003	

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year	
Other area characteristics			
Resident Population	U.S. Census Bureau, ''Table A-1: Interim Projections of the Total Population for the United States and States: April 1,2000 to July 1, 2030''.	2005	
Resident Population, Percent Change, 2000-2004	U.S. Census Bureau, *Current Population Reports*, P25-1106; ''Table 2 - Cumulative Estimates of Population Change for the United States and States, and for Puerto Rico and State Rankings: April 1, 2000 to July 1, 2004 (NST-EST2004-02)''.	2004	
Percent of the Civilian Population 18 Years and Over Who are Veterans	ACS	2003	*
Infant Mortality Rate	U.S. National Center for Health Statistics, *Vital Statistics of the United States,* annual; and unpublished data.	2002	*
Women 15 to 50 Years Old Who Had a Birth in the Past 12 Months (Per 1,000 15 to 50 years old women)	ACS	2003	*
Physicians Per 100,000 Population	American Medical Association, Chicago, IL, *Physician Characteristics and Distribution in the U.S.,* annual.	2003	
Violent Crime Rate Per 100,000 Population	U.S. Federal Bureau of Investigation, *Crime in the United States*, annual.	2003	*

See notes at end of table.

Table A-2. Listing of state-level predictors considered in the variable selection process: 2003 — Continued

State characteristics	Source	Year	
Other area characteristics (Continued)			
Federal Aid to State and Local Governments Per Capita	U.S. Census Bureau, *Federal Aid to States for Fiscal Year 2003* (issued September 2004).	2003	
State Government General Revenue Per Capita	U.S. Census Bureau; *State and Local Government Finance Estimates by State,* annual, and unpublished data.	2003	
Gross State Product in Current Dollars	U.S. BEA, Survey of Current Business, July 2005	2003	*
Energy Consumption Per Person	U.S. Energy Information Administration, *State Energy Data Report, 2001*.	2001	
Traffic Fatalities Per 100 Million Vehicle Miles, 2003	U.S. National Highway Safety Traffic Administration, *Traffic Safety Facts,* annual.	2003	
Health Care Coverage Rate for Adult	BRFSS	2003	*
Adults Aged 65+ Who Have Had a Flu Shot Within the Past Year	BRFSS	2003	*

* Indicates variables that were downloaded or key-entered for the variable selection process.

NOTE: The acronyms are ACS = American Community Survey; BEA = Bureau of Economic Analysis; BLS = Bureau of Labor Statistics; BRFSS = Behavioral Risk Surveillance System; CPS = Current Population Survey; HVS = Housing Vacancy Survey; IPEDS = The Integrated Postsecondary Education Data System; OVAE = Office of Vocational and Adult Education; and SAIPE = Small Area Income and Poverty Estimates Program.

SOURCE: U.S. Department of Commerce, Census Bureau, Census 2000 Summary File 3; U.S. Department of Commerce, Census Bureau, American Community Survey (2003); U.S. Department of Agriculture, Economic Research Service (2000); Centers for Disease Control Behavioral Risk Factor Surveillance System (2003); National Center for Health Statistics Vital Statistics of the United States (2002); Bureau of Economic Analysis Survey of Current Business (2005); U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Integrated Postsecondary Education Data System (2003); U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

This page left intentionally blank.

Appendix B

This page left intentionally blank.

Table B-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 2003

State	FIPS code[1]	Population size[2]	Percent lacking *Basic* prose literacy skills[3]	95 percent Credible interval[4] Lower bound	Upper bound
Alabama	01	3,400,000	15	11.8	19.4
Alaska	02	461,000	9	6.1	13.3
Arizona	04	4,080,000	13	9.6	18.1
Arkansas	05	2,040,000	14	10.2	17.2
California	06	26,030,000	23	20.3	26.2
Colorado	08	3,390,000	10	7.1	12.9
Connecticut	09	2,670,000	9	5.5	12.5
Delaware	10	619,000	11	6.6	16.4
District of Columbia	11	426,000	19	9.3	33.1
Florida	12	13,040,000	20	17.0	22.9
Georgia	13	6,366,000	17	14.0	20.7
Hawaii	15	944,000	16	11.5	22.2
Idaho	16	1,000,000	11	8.0	13.8
Illinois	17	9,510,000	13	10.4	16.6
Indiana	18	4,630,000	8	6.1	10.3
Iowa	19	2,250,000	7	5.3	10.1
Kansas	20	2,050,000	8	5.9	10.2
Kentucky[5]	21	3,200,000	12	10.3	14.3
Louisiana	22	3,310,000	16	12.5	20.3
Maine	23	1,040,000	7	5.2	10.2
Maryland[5]	24	4,190,000	11	9.1	13.7
Massachusetts[5]	25	5,100,000	10	8.3	12.1
Michigan	26	7,630,000	8	6.2	11.0
Minnesota	27	3,900,000	6	4.1	8.0
Mississippi	28	2,120,000	16	11.9	20.8
Missouri[5]	29	4,320,000	7	5.9	9.2
Montana	30	704,000	9	5.9	12.2
Nebraska	31	1,310,000	7	5.3	9.7
Nevada	32	1,670,000	16	9.5	25.3
New Hampshire	33	995,000	6	4.0	8.2

See notes at end of table.

Table B-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 2003—Continued

State	FIPS code[1]	Population size[2]	Percent lacking *Basic* prose literacy skills[3]	95 percent Credible interval[4]	
				Lower bound	Upper bound
New Jersey	34	6,610,000	17	13.5	20.8
New Mexico	35	1,390,000	16	12.2	21.6
New York[5]	36	15,060,000	22	19.7	25.0
North Carolina	37	6,280,000	14	11.0	16.5
North Dakota	38	489,000	6	4.2	9.0
Ohio	39	8,720,000	9	7.2	12.0
Oklahoma[5]	40	2,700,000	12	10.4	14.5
Oregon	41	2,710,000	10	7.3	13.9
Pennsylvania	42	9,560,000	13	10.2	15.5
Rhode Island	44	832, 000	8	4.7	13.9
South Carolina	45	3,100,000	15	11.6	18.4
South Dakota	46	572,000	7	4.7	9.7
Tennessee	47	4,440,000	13	10.5	16.5
Texas	48	15,940,000	19	16.4	22.1
Utah	49	1,640,000	9	6.1	13.9
Vermont	50	485,000	7	4.4	9.4
Virginia	51	5,520,000	12	9.6	14.8
Washington	53	4,640,000	10	7.3	12.8
West Virginia	54	1,420,000	13	10.2	17.2
Wisconsin	55	4,190,000	7	5.1	9.9
Wyoming	56	382,000	9	6.2	12.2

[1] The state Federal Information Processing Standards (FIPS) codes are standardized unique state identifiers. For more information about FIPS codes, see http://www.census.gov/geo/www/fips/fips.html.

[2] Estimated population size of persons 16 years and older in households in 2003.

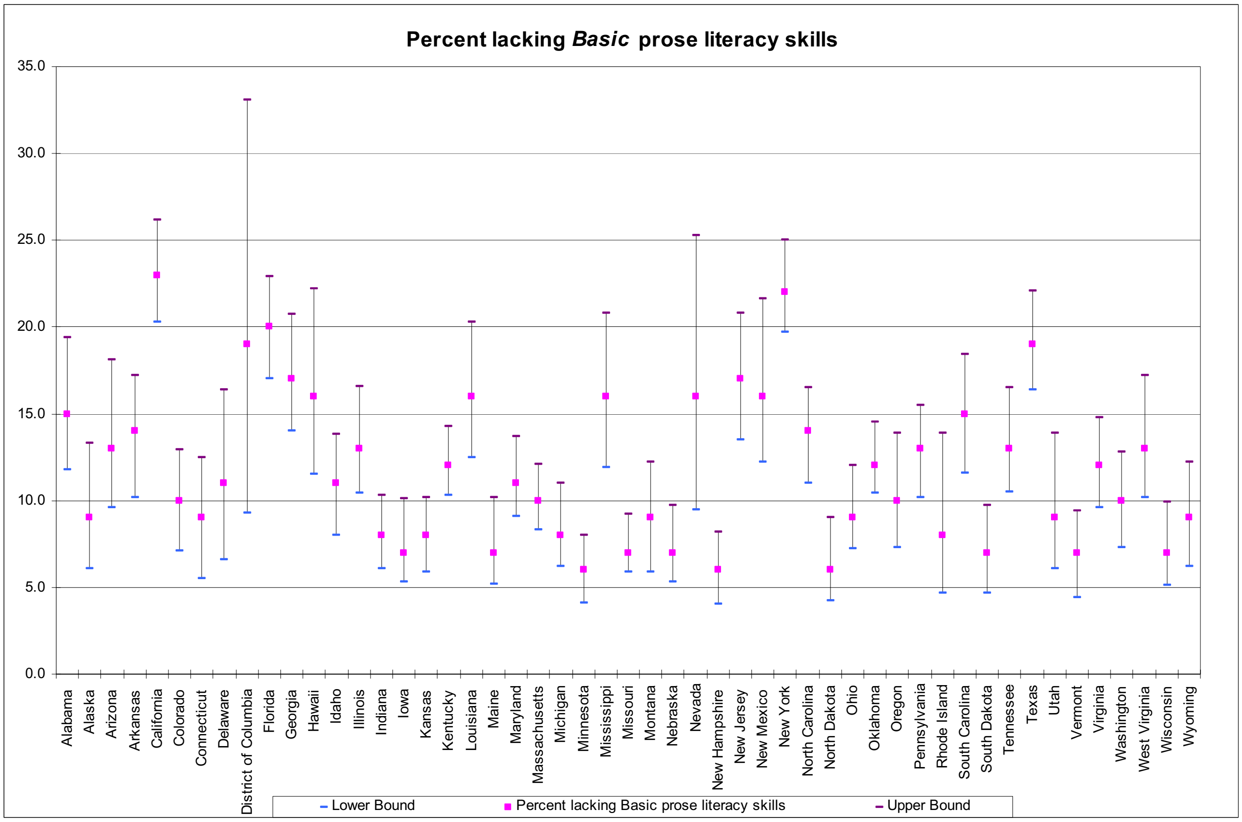
[3] Those lacking *Basic* prose literacy skills include those who could not be tested due to language barriers and those who scored below the *Basic* level in prose.

[4] The estimated percent lacking *Basic* prose literacy skills is subject to uncertainty, as measured by the associated credible interval. The probability that the true value is contained between the lower and upper bound is .95.

[5] States that paid the cost of additional assessments to obtain state-level representation. This participation is referred to as the State Assessment of Adult Literacy.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Figure B-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 2003



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

This page left intentionally blank.

Appendix C

This page left intentionally blank.

Table C-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 1992

State	FIPS code[1]	Population size[2]	Percent lacking *Basic* prose literacy skills[3]	95 percent Credible interval[4] Lower bound	Upper bound
Alabama	01	3,190,000	21	14.5	27.4
Alaska	02	416,000	10	6.0	14.0
Arizona	04	2,950,000	13	9.4	17.5
Arkansas	05	1,840,000	19	12.9	25.8
California[5]	06	23,230,000	15	11.8	17.9
Colorado	08	2,650,000	9	5.7	13.0
Connecticut	09	2,590,000	14	8.3	20.1
Delaware	10	536,000	12	7.8	15.8
District of Columbia	11	488,000	21	14.6	28.6
Florida	12	10,800,000	15	10.9	20.2
Georgia	13	5,10,000	18	12.8	24.8
Hawaii	15	889,000	18	13.9	23.2
Idaho	16	779,000	10	6.4	13.9
Illinois[5]	17	8,930,000	15	12.3	18.2
Indiana[5]	18	4,350,000	10	7.4	14.0
Iowa[5]	19	2,160,000	7	4.3	9.9
Kansas	20	1,910,000	9	5.7	13.0
Kentucky	21	2,900,000	19	13.2	26.3
Louisiana[5]	22	3,170,000	21	15.4	27.1
Maine	23	957,000	13	7.4	18.8
Maryland	24	3,790,000	12	8.0	17.2
Massachusetts	25	4,760,000	13	8.7	17.8
Michigan	26	7,200,000	12	8.5	16.2
Minnesota	27	3,390,000	9	5.4	12.1
Mississippi	28	1,950,000	25	17.9	34.0
Missouri	29	3,990,000	13	8.5	17.6
Montana	30	617,000	9	5.7	13.1
Nebraska	31	1,220,000	8	5.3	12.3
Nevada	32	1,040,000	13	9.7	17.7
New Hampshire	33	855,000	11	6.4	16.3

See notes at end of table.

Table C-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 1992—Continued

State	FIPS code[1]	Population size[2]	Percent lacking *Basic* prose literacy skills[3]	95 percent Credible interval[4]	
				Lower bound	Upper bound
New Jersey[5]	34	6,160,000	16	12.2	19.6
New Mexico	35	1,170,000	17	11.1	24.3
New York[5]	36	14,190,000	16	12.9	20.1
North Carolina	37	5,380,000	18	12.6	24.6
North Dakota	38	482,000	11	7.2	16.2
Ohio[5]	39	8,450,000	12	8.5	15.4
Oklahoma	40	2,440,000	13	8.5	18.8
Oregon	41	2,300,000	10	6.2	13.3
Pennsylvania[5]	42	9,440,000	13	9.8	17.3
Rhode Island	44	799,000	18	12.7	23.4
South Carolina	45	2,760,000	20	14.0	28.1
South Dakota	46	526,000	11	6.7	15.2
Tennessee	47	3,910,000	19	13.0	25.3
Texas[5]	48	13,110,000	18	13.5	22.7
Utah	49	1,250,000	8	5.2	12.4
Vermont	50	439,000	11	6.4	15.9
Virginia	51	4,970,000	15	10.2	20.7
Washington[5]	53	3,920,000	7	4.9	10.0
West Virginia	54	1,420,000	17	11.6	24.2
Wisconsin	55	3,820,000	10	6.1	13.6
Wyoming	56	342, 000	9	5.3	12.4

[1] The state Federal Information Processing Standards (FIPS) codes are standardized unique state identifiers. For more information about FIPS codes, see http://www.census.gov/geo/www/fips/fips.html.

[2] Estimated population size of persons 16 years and older in households in 1992.

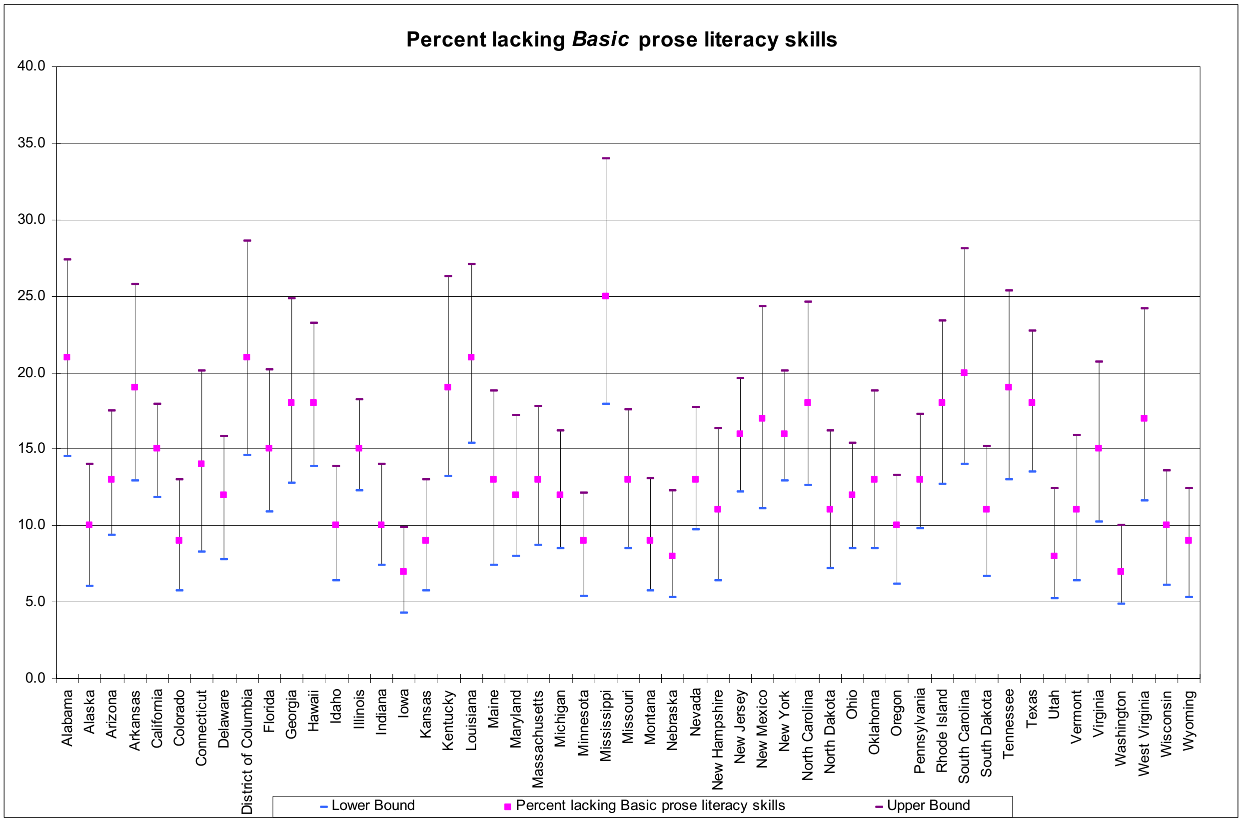
[3] Those lacking *Basic* prose literacy skills include those who could not be tested due to language barriers and those who scored below the *Basic* level in prose.

[4] The estimated percent lacking *Basic* prose literacy skills is subject to uncertainty, as measured by the associated credible interval. The probability that the true value is contained between the lower and upper bound is .95.

[5] States that paid the cost of additional assessments to obtain state-level representation. This participation is referred to as the State Adult Literacy Survey.

SOURCE: U.S. Department of Education, National Center for Education Statistics, 1992 National Adult Literacy Survey.

Figure C-1. Indirect estimates of the percent lacking *Basic* prose literacy skills and corresponding credible intervals, by state: 1992



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey.