CHAPTER 12

WEIGHTING AND VARIANCE ESTIMATION

Thomas Krenzke and Leyla Mohadjer, Westat

The National Assessment of Adult Literacy (NAAL) sample includes both a household component and a prison component. The household component includes two sets of household samples: (1) a national NAAL household sample and (2) household samples from six states, used to administer the State Assessment of Adult Literacy (SAAL). A prison component, involving a sample of adult inmates in federal and state prisons, was conducted to improve the representation of the target population. The complex sample design involved variable sampling rates, stratification, and several stages of selection. To make valid inferences from the responding adults to the target population, the sample must be weighted to account for the special sample design features as well as other complexities arising from nonresponse. In addition, simple formulas (that assume simple random sampling) for variance estimation are not appropriate. Even if sampling weights are used to construct the survey estimates, inferences will not be valid unless the corresponding variance estimator appropriately reflects all the complex features of the sample design. The complex weighting procedures were used to combine the national and state household samples and the prison samples, account for oversampling, and reduce the bias due to nonresponse.

This chapter is divided into two major subsections. The first, section 12.1, discusses the weighting and variance estimation procedures for the NAAL and SAAL household samples. The second, section 12.2, describes the weighting and variance estimation procedures for the correctional institution sample, referred to here as the prison study sample.

12.1 HOUSEHOLD SAMPLES

Differential probabilities of selection were adjusted by computing base weights for all adults selected into the household samples. The base weight was calculated as the reciprocal of a respondent's final probability of selection. Further, to combine the NAAL and SAAL household samples, composite weights were calculated for the respondents in the six participating states and the respondents in the national NAAL household sample located within the six SAAL states. Finally, to adjust for nonresponse, weights were adjusted through poststratification and raking to match the 2003 Current Population Survey (CPS) data. The remainder of this section provides detailed information on the weighting and variance estimation procedures used for the household samples.

This section begins by describing the preliminary steps in weighting the household samples (section 12.1.1). The first steps involved computing base weights and nonresponse adjustments for the dwelling units selected for screening (section 12.1.2). Once the screener weighting steps were processed for the NAAL sample and each of the six SAAL states, weighting steps began for the sample persons selected for the background questionnaire. The background questionnaire weighting steps were done separately for each NAAL and SAAL household sample and involved computing base weights, making nonresponse adjustments, and trimming the weights (section 12.1.3). Before compositing, the household sample weights were calibrated to known population estimates (section 12.1.4). After calibration, in order to combine the NAAL and SAAL household samples into one sample, the weights were composited (section 12.1.5). The composited weights were adjusted using a raking procedure, as described in section 12.1.6. Finally, replicate weights were created using the stratified jackknife method, as described in section 12.1.7.

Sample weights were produced for sample persons who either completed the background questionnaire or could not complete the background questionnaire owing to language problems or mental disabilities. The purpose of calculating sample weights was to permit inferences from sample persons to the populations from which they were drawn and to have the tabulations reflect estimates of the population totals. The sample weighting process was designed to accomplish the following objectives:

1. Permit unbiased estimates, taking into account the fact that all persons in the population did not have the same probability of selection;
2. Minimize the potential bias arising from differences between respondents and nonrespondents;
3. Combine the state and national samples in an efficient manner;
4. Use auxiliary data on known population characteristics in such a way as to reduce sampling errors and to bring data up to the dimensions of the population totals;
5. Reduce the variation of the weights and prevent a small number of observations from dominating domain estimates; and
6. Facilitate sampling error estimation under complex sample designs.

Objective 1 was accomplished by computing base weights for the households selected for screening and, subsequently, for persons selected for the background questionnaire and assessment from the eligible participating households. The details of the base weight calculations for the screener and the background questionnaire are presented in sections 12.1.2.1 and 12.1.3.1, respectively.

Objective 2 was accomplished through nonresponse weighting adjustments that accounted for screener nonresponse and background questionnaire nonresponse. Sections 12.1.2.2 and 12.1.3.2 discuss the nonresponse adjustments for the screener and background questionnaire, respectively. Some reduction in potential bias was also achieved while meeting Objective 4 by calibrating the weights. This was accomplished by using weighting variables that were not used for nonresponse adjustment because data were available only for respondents.

Objective 3 was addressed through the composite weighting procedure. Composite weights were computed for the respondents in the six state samples and the respondents in the national sample primary sampling units (PSUs) of those six states. Area sampling procedures that included stratification, PSU formation, sample design, and selection at the various stages of sampling were applied to the national and state components. Further, the same instruments were used to screen households and to collect background information and literacy assessment data in the state and national surveys. To take full advantage of this comparability, the samples were combined to produce both state- and national-level statistics. Section 12.1.5 describes the composite estimation procedures.

To meet Objective 4, the weights were calibrated to known totals from the 2003 March Supplement of the CPS.[1] The weights were raked so that numerous totals calculated with the resulting full-sample weights would agree with the CPS totals. Calibration procedures were implemented for both the national sample within each state and the state sample prior to compositing the weights. After the weights had been composited, another raking process was conducted to rescale the weights. The calibration procedures are described in sections 12.1.4 and 12.1.6.2.

Objective 5 was addressed by trimming the weights. A small number of weights were reduced using a type of inspection approach (referred to as the k x median rule) within prespecified sampling and analytical domains. The trimming procedure was implemented twice during the weighting process, once before compositing the weights and once after compositing. For more discussion of the trimming procedure, refer to sections 12.1.3.3 and 12.1.6.2.2.

Finally, Objective 6 was accomplished by creating 61 replicate weights using the stratified jackknife method. The NCES standards ask for the number of replicates to be greater than 29 and less than 101. Full-sample and replicate weights were calculated for each record to facilitate the computation

[1] The March CPS supplement is an annual survey, conducted by the Bureau of Labor Statistics and the U.S. Census Bureau, to collect detailed information on demographics, income, and work experience.

of unbiased estimates and their standard errors. The weighting procedures were repeated for 61 strategically constructed subsets of the sample to create a set of replicate weights for variance estimation using the jackknife method. The replication scheme was designed to produce stable estimates of standard errors for the national and six individual state estimates. The replication design and the significance of the number of replicates is discussed further in section 12.1.7. The variance strata and variance units created for the replication process can also be used in estimating sampling error using Taylor series approximation (Wolter 1985).

Prior to the weighting process, it was necessary to resolve any issues related to the data used in weighting. The next section discusses this preliminary data cleaning procedure.

12.1.1 Preliminary Steps in Weighting

The data used in the weighting process underwent consistency checks to prevent any errors in the sample weights. The checks were performed only on variables required for weighting and were limited to records that required weights.

The consistency checks also helped identify any unusual values. Westat prepared listings of records with missing values in any of the weighting variables. The listings showed the following variables: the respondent's case identification (ID) number, age, date of birth, gender, race/ethnicity, country of birth, and level of education; race of the head of household; and the number of age-eligible members and respondents in the household. The printed listings were used to review the extent of missing data, identify the pattern of missing data, and prepare for imputation. The age, gender, and race/ethnicity data from the screener and the background questionnaire were also compared for consistency. Inconsistency in survey data is due to individuals reporting data for others in the screener. In all, less than 1 percent had missing data or inconsistent data between the screener and background questionnaire for these items.

The weighting variables that were at a finer level of detail than was necessary for the later steps of weighting (age, gender, race/ethnicity, country of birth, and level of education) were recoded (i.e., collapsed to the required levels). Age, race/ethnicity, and gender were collected in both the screener and the background questionnaire, thereby providing two measures of the same item. The background questionnaire measure was preferred for all items. For the few cases in which the background questionnaire measure was missing, the screener value was used as a direct substitute.

For level of education and country of birth, which were not collected through the screener, a limited amount of imputation was performed to fill in the data for respondents so that the variables could be used in the calibration and raking processes. To the extent possible, missing values were filled in using information from other items in the background questionnaire. For example, if the data on country of birth were missing, the questions regarding length of residence in the United States, education obtained before coming to the United States, and language the respondent first learned to read or write were consulted to determine whether the respondent was born in the United States. Similarly, several education variables were used to create a three-level education measure (less than high school, high school or equivalent, or more than high school) if this variable had missing data.

If no other background questionnaire data were available for imputing these items, and since there were a small number of missing values remaining, a simple imputation procedure was performed as follows.[2] Two cases still had missing values for the "born in the U.S." item. To obtain values for these cases, cells were formed by PSU and segment. Then the most frequent value in the cell was given to the missing case (i.e., modal within cell hotdeck[3]). For the seven remaining cases with missing education data, cells were formed by PSU, age (16–19, 20–29, 30–69, 70+), and race/ethnicity. Again, the most frequent value for education in the cell was given to the missing case.

Some additional dwelling units came into the sample as a result of the missed structure and hidden dwelling unit procedures (refer to section 7.1.3.5 for more information), which allowed units that were missed in the segment listing activities to be included in the sample with a known probability of selection. All newly discovered dwelling units within a segment were included unless the total number was unusually large, in which case a sample of newly discovered dwelling units was taken. Whenever a sample of missed units was selected, detailed records indicated the PSU, segment, number of new dwelling units selected, and total number of newly discovered dwelling units. This information was attached to each of these records prior to the calculation of base weights.

A few final checks were run (refer to section 12.1.6.3 for further discussion) before the screener base weights were calculated to ensure the availability and validity of all fields required by the base weights program (fields created for the special cases mentioned above and fields for the total number of age-eligible household members and the number of sample persons for each dwelling unit). A detailed description of the screener base weight computation is provided in the next section.

[2] For the 355 nonrespondents who did not complete the survey because of language problems or mental disability, the imputation method was more complex. Details are provided in section 12.1.3.2.2.

[3] Hotdeck is an imputation procedure that uses data from the same sample survey.

12.1.2 Screener Base Weights and Nonresponse Adjustments

To produce unbiased estimates, differential weights must be used for various subsets of the population whenever subsets have been sampled at different rates. Weighting was required to account for the oversampling of Blacks and Hispanics in high-minority segments of the national sample, as discussed in section 7.1.3.3. The screener data helped determine the probabilities of selection for the screener. Section 12.1.2.1 summarizes the base weight computation for the household samples.

If every selected household had agreed to complete the screener and every selected person had agreed to complete the background questionnaire and the assessment booklet, weighted estimates based on the data would be approximately unbiased (from a sampling point of view). However, nonresponse occurs in any survey operation, even when participation is mandatory, and adjustments are always necessary to avoid potential nonresponse bias. The weighting adjustments for screener nonresponse are discussed in section 12.1.2.2.

12.1.2.1 Screener Base Weights

The probability of a dwelling unit k being selected into NAAL or SAAL, denoted as Pijk(mdu) (as given in table 7-8), is the product of the conditional probabilities at the PSU, segment, and dwelling unit levels. Other factors entering into the probability of selection were due to chunking (refer to section 7.1.2.3), dwelling unit selection from segments selected for both NAAL and SAAL (refer to section 7.1.3.3), missed dwelling units identified through the missed structure process (refer to section 7.1.3.5), and subsampling of nonminority dwelling units in oversampled high-minority segments (refer to section 7.1.3.3). The screener base weights were computed as the reciprocal of the probability of selection of dwelling unit k of PSU i and segment j, after accounting for subsampling due to the missed dwelling units (mdu) procedure, as shown in the following formula:

$$W_{ijk}^{base,SCR} = \frac{1}{P_{ijk(mdu)}}.$$

Table 12-1 shows the distribution of the screener base weights for the NAAL sample and for each of the SAAL samples. The variation—as seen by the minimum, maximum, and coefficient of variation—can be explained by several factors. These factors include oversampling of Blacks and Hispanics, sampling of missed dwelling units in segments where a large number of dwelling units were found by the

interviewers as they canvassed the listing area, and a small number of unique sampling situations. The table also indicates that the coefficient of variation is much lower for the SAAL states than for the national NAAL sample due to an equal probability design for households.

Table 12-1. Screener base weight distribution for the household samples, by sample: 2003

Sample	Screener base weights				
	Sample cases	Median	Minimum	Maximum	Coefficient of variation (percent)[1]
NAAL	25,450	7,240	1,207	21,719	45
SAAL					
Kentucky	2,306	771	386	1,542	6
Maryland	1,493	1,528	764	2,292	11
Massachusetts	1,509	1,750	875	3,499	8
Missouri	1,499	1,658	829	2,487	5
New York	1,499	5,151	2,575	5,151	2
Oklahoma	1,609	992	496	2,975	27

[1] The coefficient of variation is the standard deviation of the weights divided by the mean weight.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

12.1.2.2 Screener Nonresponse Adjustment

For the screener nonresponse adjustment, the nonrespondents were divided into two categories. The first category consisted of cases involving nonliteracy-related nonresponse, such as refusals and nonresponse because of illness. Nonliteracy-related nonrespondents were likely to be similar to respondents with respect to English literacy scores. The second category was literacy-related nonresponse. Language problems are the only type of literacy-related nonresponse at the screener level, with only 160 such cases in the NAAL and SAAL household samples. Households with this type of nonresponse were presumed to differ from responding households with respect to literacy. Therefore, the weighting procedures adjusted the weights of the respondents to represent the nonliteracy-related nonresponse only. The weights of the language problem cases were not adjusted during the screener-level nonresponse adjustment because their literacy status was expected to differ from that of respondents. The contribution of the screener-level literacy-related nonresponse to the total population was accounted for by literacy-related nonresponse adjustment carried out for the background questionnaire sample (refer to section 12.1.3.2.2).

Little was known about the nonresponding households, including their eligibility.[4] Before any nonresponse adjustment was processed, an adjustment for unknown eligibility was performed. In this step, the weights of the households with unknown eligibility status, such as those with maximum callbacks, were distributed among the cases with known eligibility status. The second step distributed the weights of the eligible nonrespondents among the eligible respondents.

All adjustments were made within weighting classes. Because very little was known about the households that did not respond to the screener, information used to form weighting classes had to come from a different source. The frame contained only aggregate demographic information, such as region and Metropolitan Statistical Area (MSA) status. However, because the sampling was performed using census geography, the sampled segments were merged to the Census 2000 Summary File 3 (SF3)[5] files to create segment-level weighting variables by extracting segment-level census data.

Prior to the weight adjustments, classification software was used to help identify weighting classes for the adjustments for unknown eligibility status and nonresponse. A Chi-squared Automatic Interaction Detector (CHAID) (Kass 1980) was used to help identify important variables to be used in forming weighting classes that were homogeneous in terms of response propensity. CHAID is a classification algorithm that divides a population into homogeneous subgroups with respect to a target characteristic (the dependent variable). Once the weighting variables were identified through CHAID, the weighting classes were formed through a hierarchical ordering of the weighting variables. The hierarchical ordering was formed using the general order as they were selected for the CHAID tree classification. Table 12-2 shows the variables selected to form the weighting classes for the NAAL and SAAL household screener samples based on the CHAID results. Weighting classes were combined if the cell size was less than 30 or the adjustment factor was greater than 1.50. The criteria for cell size and maximum adjustment factor is a guideline and can vary from survey to survey, and by weighting stage within a survey (Kalton and Kasprzyk 1986).

[4] Households were ineligible only if they were vacant or were not a residential dwelling unit.

[5] The SF3 files contain data from the 52-item census long form that was issued to about 19 million households. The files contain data on demographics, education, income, commuting, and other characteristics.

For each weighting cell α, the screener unknown eligibility adjustment factor is computed as follows:

$$F_{\alpha}^{unk,SCR} = \frac{\sum_{k \in S(\alpha)} W_{ijk}^{base,SCR}}{\sum_{k \in SK(\alpha)} W_{ijk}^{base,SCR}},$$

where

S(α) = the set of sampled cases (i.e., STATUS = 0, 1, 2, 3, or 4) in weighting cell *α* and

SK(α) = the set of sampled cases with known eligibility status (i.e., STATUS = 0, 1, 2, or 3) in weighting cell *α*

and where

STATUS = 0 literacy-related nonrespondents (language problems only);

1 respondents;

2 nonrespondents known to be eligible, including respondents who refused and those unavailable due to illness;

3 ineligibles, including households subsampled out (those with nonminority reference persons in high-minority segments), vacancies, and sampled cases that were not dwelling units; and

4 cases for which the eligibility status was not known.

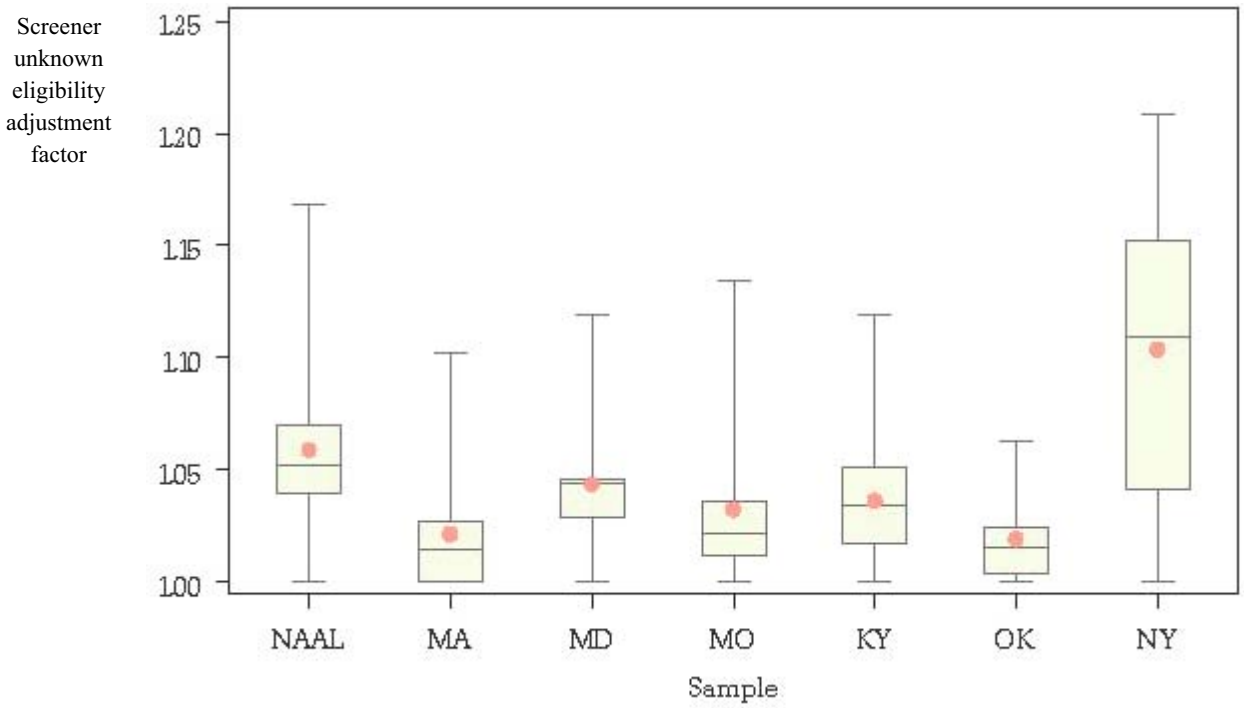
Table 12-2. Variables used in forming weighting classes for the screener nonresponse adjustment, by sample: 2003

Sample	Variables
NAAL	Indicator that percentage of Black or Hispanic population in segment exceeds 12.5 percent
	Percentage of segment population who do not speak English at home but speak English well
	Percentage of segment population below 150 percent of poverty
	Census region
SAAL	
Kentucky	Median household income in segment
	Percentage of segment population who speak Spanish at home but do not speak English well or at all
	Percentage of segment population with less than a high school education
Maryland	Percentage of segment population with a high school education
	Percentage of segment population who speak a language other than Spanish or English at home but speak English well
	MSA status of PSU
	Percentage of segment population below 150 percent of poverty
Massachusetts	Indicator that percentage of Black or Hispanic population in segment exceeds 12.5 percent
	Percentage of segment population with more than a high school education but less than a bachelor's degree
	Median household income in segment
Missouri	Whether the segment is in an urban area
	Median household income in segment
	Percentage of segment population with a bachelor's degree
New York	Percentage of segment population who speak English only
	Percentage of segment population with a bachelor's degree or higher

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

The distribution of the adjustment factors for cases with known eligibility in the NAAL and SAAL samples is shown in figure 12-1. The figure displays a box-and-whisker plot that shows the median (the horizontal line inside the box), the mean (the dot inside the box), the 25th and 75th percentiles (bottom and top of box, repectively), and the minimum and maximum values (the end of the line below the box, and the end of the line above the box, respectively). The figure shows that the adjustment factors for the national NAAL sample range from 1.0 to about 1.7. New York's SAAL sample has the largest range (from 1.0 to 1.2). The other SAAL states' average adjustment factors are small (less than 1.05).

Figure 12-1. Distribution of the unknown eligibility adjustment factors for the household samples, by sample: 2003



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

Subsequently, the screener nonresponse adjustment factor was computed in the following way:

$$F_{\alpha}^{nr,SCR} = \frac{\sum_{k \in SE(\alpha)} W_{ijk}^{base,SCR} F_{\alpha}^{unk,SCR}}{\sum_{k \in SC(\alpha)} W_{ijk}^{base,SCR} F_{\alpha}^{unk,SCR}},$$

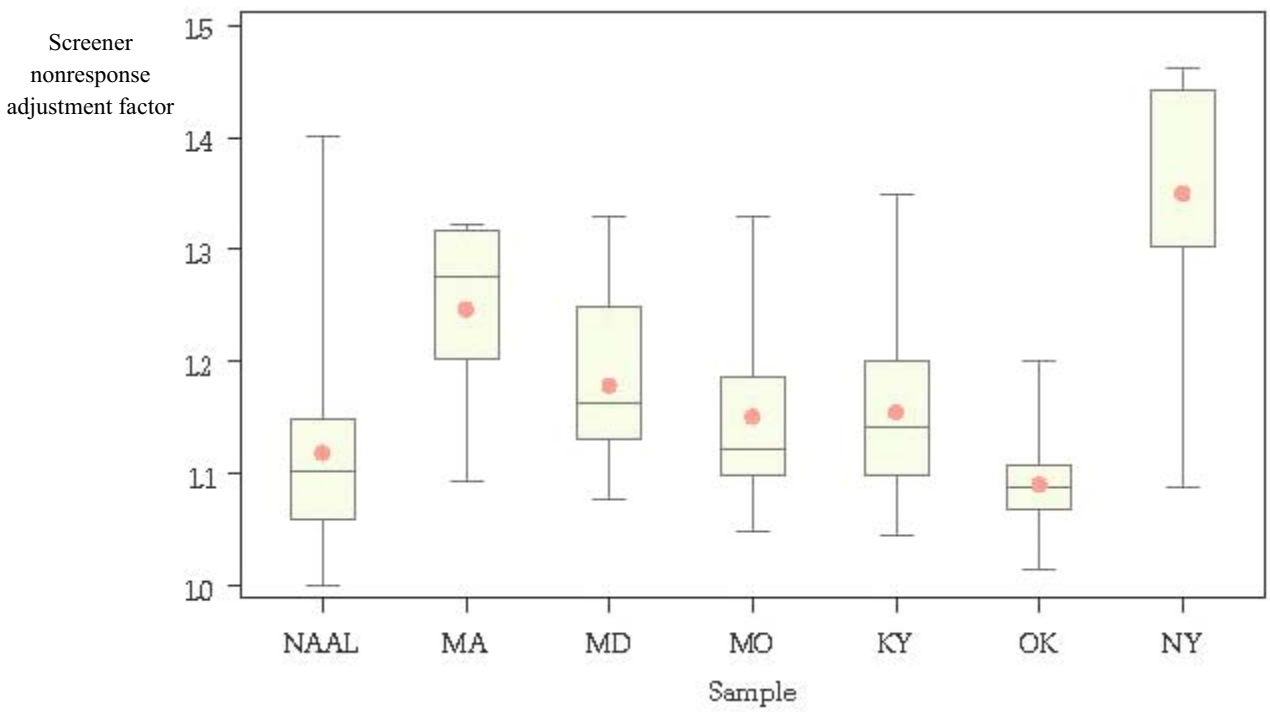
where

SE(α) = *the set of eligible sampled cases (i.e., STATUS category 1 or 2) in weighting cell α and*

SC(α) = *the set of completed cases (i.e., STATUS category 1) in weighting cell α.*

For simplicity, the notation assumes the same cells as used in the unknown eligibility adjustment, when in fact the cell definitions changed as a result of cell collapsing when the number of respondents was less than 30 or adjustment factors were greater than 1.50. The distribution of the screener nonresponse adjustment factors for screener respondents in the household samples is shown in figure 12-2. The figure shows that New York's SAAL sample had relatively high adjustment factors on average due to its relatively low screener response rates. The national NAAL sample's adjustment factors range from 1.0 to about 1.4.

Figure 12-2. Distribution of the screener nonresponse adjustment factors for the household samples, by sample: 2003



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

The adjustment $F_{\alpha}^{nr,SCR}$ was applied only to the unknown eligibility-adjusted weights of the screener completes (i.e., STATUS category 1). That is, the nonresponse-adjusted weight, $W_{ijk}^{nr,SCR}$, was computed as follows:

$W_{ijk}^{nr,SCR} = W_{ijk}^{base,SCR} F_{\alpha}^{unk,SCR}$, if dwelling unit k was a literacy-related nonrespondent (STATUS category 0);

= $W_{ijk}^{base,SCR} F_{\alpha}^{unk,SCR} F_{\alpha}^{nr,SCR}$, if dwelling unit k was a respondent (STATUS category 1); and

= 0, if dwelling unit k was a nonrespondent, ineligible, or of unknown eligibility status (STATUS category 2, 3, or 4).

12.1.3 Background Questionnaire Base Weights, Nonresponse Adjustments, and Trimming

The derivation of base weights was necessary to prevent potentially serious biases in the outcome statistics. The study specifications called for the selection of one person in households with fewer than four eligible members and two persons in households with four or more eligible members. Members of households with only one eligible member had twice the chance of selection as those in households with two (or four) eligible members. To produce unbiased estimates, different weights had to be used to account for the within-household selection rate. Section 12.1.3.1 summarizes the base weight computation for the background questionnaire sample, section 12.1.3.2 presents the background questionnaire nonresponse adjustment procedures, and section 12.1.3.3 describes the trimming procedure used to reduce the impact of extreme weights.

12.1.3.1 Background Questionnaire Base Weights

The background questionnaire base weights were computed as the product of the screener nonresponse-adjusted weight and the reciprocal of the within-household probability of selection for person l within household k of PSU i and segment j, as shown in the following formula:

$$W_{ijkl}^{base,BQ} = W_{ijkl}^{nr,SCR} \frac{1}{CP_{ijkl}},$$

where

CP_{ijkl} = the within-household probability of person l being selected into NAAL or SAAL, which is the ratio of the number of persons selected in household k to the number of eligible persons in household k.

Table 12-3 shows the distribution of the background questionnaire base weights for NAAL and for each of the SAAL household samples. The oversampling of Blacks and Hispanics resulted in a larger coefficient of variation for the national NAAL household sample, as expected. Other major reasons for the variation in the sampling weights include sampling of missed dwelling units in segments where a large number of dwelling units were discovered by the interviewers as they canvassed the listing area, the number of eligible persons in the household, and the screener nonresponse adjustment factors.

Table 12-3. Distribution of the background questionnaire base weights for household samples, by sample: 2003

Sample	Number of sample persons	Median	Minimum	Maximum	Coefficient of variation (percent)
NAAL	16,409	8,797.33	1,283.36	87,353.59	63.09
SAAL					
Kentucky	1,694	1,766.13	815.15	3,301.97	35.58
Maryland	1,290	3,719.97	1,652.13	8,369.14	37.26
Massachusetts	1,116	4,495.24	1,070.26	13,714.16	33.69
Missouri	1,368	3,703.51	1,739.19	7,826.35	38.35
New York	956	14,851.10	3,824.44	26,789.43	37.71
Oklahoma	1,293	2,189.19	1,016.34	6,622.13	38.66

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

12.1.3.2 Background Questionnaire Nonresponse Adjustment

12.1.3.2.1 Nonliteracy-Related Nonresponse

At the background questionnaire level, separate adjustments were made for the literacy-related nonrespondents and the other nonrespondents. This section discusses weighting adjustments for nonliteracy-related nonresponse. For the household samples, the variables available for nonresponse adjustments for the background questionnaire included variables from the Census 2000 SF3 file and screener variables (region, age, race/ethnicity, and gender). The weighting variables used in the screener nonresponse adjustment were also considered for the background questionnaire adjustment.

The sample persons were classified into the following STATUS groups:

STATUS = 0 literacy-related nonrespondents (language problems only);

1 respondents; and

2 nonliteracy-related nonrespondents.

The classification software, CHAID, was used to identify the key variables to use to form the weighting classes. Table 12-4 shows the variables selected to form the weighting classes for the household samples based on the CHAID analysis. More discussion on this approach is given in section 12.1.2.2. After the variables were identified, the weighting classes were formed through a hierarchical ordering of the weighting variables.

Table 12-4. Variables forming background questionnaire nonresponse adjustment weighting classes, by sample: 2003

Sample	Variables	Source
NAAL	Percentage of segment population below 150 percent of poverty	SF3[1]
	Percentage of segment population who speak Spanish and speak English well	SF3
	Gender	Screener
	Age category	Screener
	Household size	Screener
SAAL		
Kentucky	Percentage of segment population with less than a high school education	SF3
	Age category	Screener
	Gender	Screener
Maryland	MSA status of PSU	Sampling files
	Age category	Screener
	Percentage of segment population who speak another language at home and speak English well	Screener
Massachusetts	Median household income in segment	SF3
	Race/ethnicity	Screener
	Percentage of segment population with more than a high school education	SF3
Missouri	Age category	Screener
	Gender	Screener
	Median household income in segment	SF3
New York	Percentage of segment population below 150 percent of poverty	SF3
	Race/ethnicity	Screener
	Percentage of segment population with more than a college education	SF3
Oklahoma	Median household income in segment	SF3
	Household size	Screener
	Percentage of segment population with more than a high school education	SF3

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

Once the weighting classes had been identified, the nonresponse adjustment factors were computed. Weighting classes were combined if the cell size was less than 30 or the adjustment factor was greater than 1.75. The maximum adjustment of 1.75 is larger than that for the screener adjustment. There is not a fixed rule for the maximum, although the statistician attempts to balance an increase in variance due to large adjustments, with decrease in bias due to nonresponse. Refer to Kalton and Kasprzyk (1986) for more discussion. The corresponding sample-based nonresponse adjustment is defined to be the ratio of sums:

$$F_{\beta}^{nr,BQ} = \frac{\sum_{l \in SN(\beta)} W_{ijkl}^{base,BQ}}{\sum_{l \in SC(\beta)} W_{ijkl}^{base,BQ}},$$

where

SN(β) = *the set of completed background questionnaires and nonliteracy-related nonrespondents (STATUS category 1 or 2) in weighting class β and*

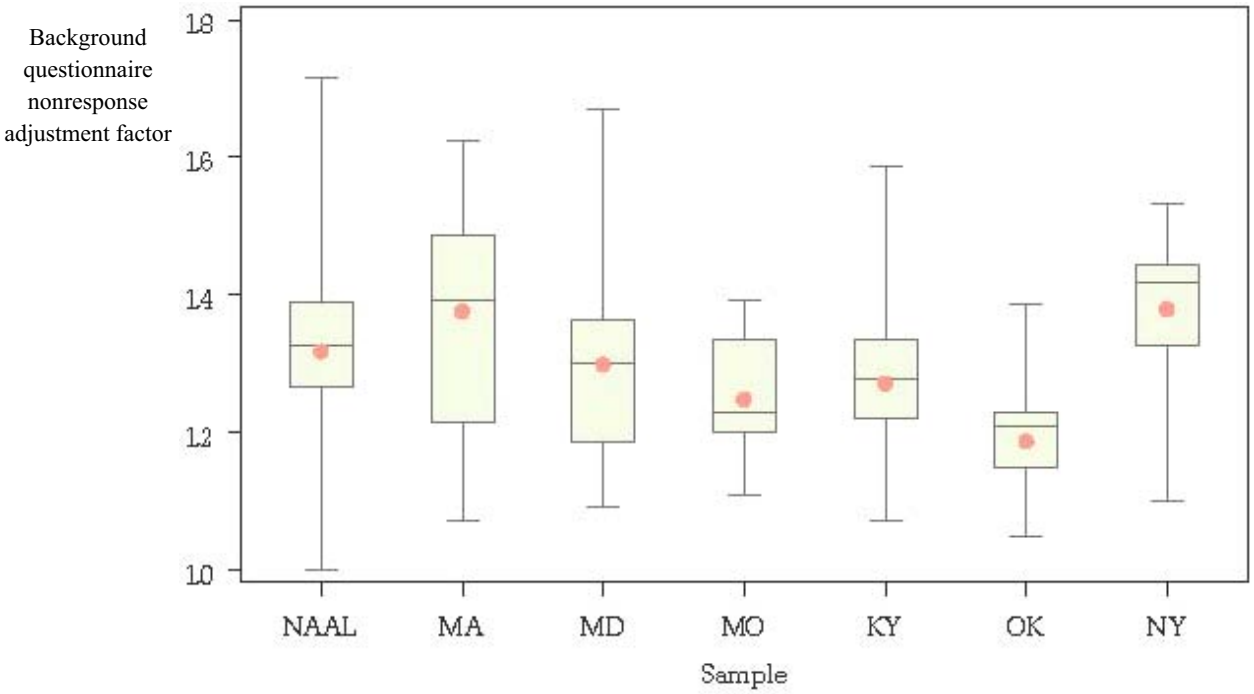
SC(β) = *the set of completed background questionnaires (STATUS category 1) in weighting class β.*

The distribution of the background questionnaire nonresponse adjustment factors for screener respondents in the NAAL and SAAL samples is shown in figure 12-3. The figure shows that the national NAAL sample's adjustment factors ranged from 1.0 to about 1.7. Oklahoma had a low adjustment factor on average, at about 1.2.

12.1.3.2.2 Literacy-Related Nonresponse

Of the 355 sample persons who did not complete the background questionnaire for literacy-related reasons, 211 sample persons had language problems and 144 sample persons had mental disabilities as determined by the interviewers and documented in the noninterview reports (refer to chapter 8 for more discussion). These cases were included in the background questionnaire data file along with their age, race, and gender information from the screener. Educational attainment and country of birth, two variables needed for calibrating the weights (section 12.1.4), were imputed using logistic regression models that included segment-level education and poverty data from the Census 2000 SF3 data.

Figure 12-3. Distribution of the background questionnaire nonresponse adjustment factors for the household samples, by sample: 2003



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

Through the logistic regression models, the predicted values (response propensities or probabilities) were generated for each classification of education (or country of birth). The imputed value was assigned through a random draw from the probability distribution, as predicted by the model. This approach is discussed in Thibaudeau et al. (1997).

Before the background questionnaire weights were calibrated, the background questionnaire literacy-related respondent weights were adjusted to account for the 160 literacy-related screener nonrespondents. This adjustment was necessary primarily to allow the literacy-related background questionnaire respondents to represent the literacy-related screener nonrespondents in the calibration procedure. This adjustment assumed that the literacy-related nonrespondents to the screener and the background questionnaire are similar in literacy. The weighting class, β, was simply the national NAAL household sample and each of the six SAAL states. The corresponding sample-based nonresponse adjustment is defined to be the ratio of sums:

$$F_{\beta}^{nr,BQ} = \frac{\sum_{k \in SL(\beta)} W_{ijk}^{base,SCR}}{\sum_{k \in S0(\beta)} W_{ijk}^{base,SCR}},$$

where

SL(β)= *the set of sample dwelling units with either a literacy-related screener nonresponse or a literacy-related background questionnaire nonrespondent in weighting class β and*

S0(β)= *the set of literacy-related background questionnaire nonrespondents (STATUS category 0) in weighting class β.*

12.1.3.3 Background Questionnaire Trimming Adjustment

A trimming algorithm was used to reduce the variation in the background questionnaire nonresponse-adjusted weights. Reasons for the variation in the NAAL and SAAL sampling weights include subsampling of newly discovered dwelling units, number of eligible persons in the household, and screener and background questionnaire nonresponse adjustment factors.

In general, trimming procedures introduce some bias into the sampling weights (Lee 1995). However, as Lee discusses, the trimming adjustment in most cases will reduce the sampling error component of the overall mean square error more than it increases the bias when the adjustment is applied to only a very small number of weights. Trimming cells were formed by crossing the high/low-minority segment indicator (defining sampling domains) with a three-category race variable (defining analytical domains). Within each cell, cases that had weights greater than three times the median were considered for having their weights reduced. (This approach is hereinafter referred to as the 3× median rule.) This type of inspection approach, which is very common in survey weighting practices, is discussed in Potter (1990). The trimming factor, denoted by $F_{ijkl}^{trim,BQ}$, was the ratio of the cutoff value to the background questionnaire nonresponse adjustment weight. The trimming factor for the full-sample weights was then applied to the replicate weights (refer to section 12.1.7 for a discussion of replicate weights as they pertain to variance estimation).

For the NAAL household sample, 52 full-sample weights were trimmed. Table 12-5 shows the number of weights trimmed in each cell and the distribution of trimmed weights and trimming factors. All 17 cases requiring trimming in the "other" race category were due to the subsampling of dwelling units

found during the missed structure procedure. For the Hispanic and Black trimmed cases, the large background questionnaire nonresponse-adjusted weights were due mainly to large nonresponse adjustment factors. For the SAAL states, no trimming was needed for Kentucky, Maryland, Missouri, New York, or Oklahoma. One case involving missed dwelling unit subsampling was trimmed for Massachusetts.

Table 12-5. Distribution of trimmed weights and trimming factors, by minority status and race: 2003

Minority status of segment	Race	No. of weights trimmed	Trimmed weight				Trimming factors	
			N	Mean	Coefficient of variation	Max	Mean	Coefficient of variation
Total		52	12,753					
Low	Hispanic	0	282	22,632.22	34.01	46,178.90	1.0000	0.00
Low	Non-Hispanic Black	0	197	19,881.10	48.39	47,995.49	1.0000	0.00
Low	Other	15	5,915	21,971.97	39.59	68,246.93	0.9996	0.91
High	Hispanic	20	2,587	7,217.94	41.25	20,578.82	0.9987	1.71
High	Non-Hispanic Black	15	2,640	6,231.19	47.07	18,582.41	0.9986	2.27
High	Other	2	1,132	17,112.82	50.89	52,981.88	0.9997	0.83

NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

12.1.4 Calibration Adjustments Prior to Compositing

Undercoverage of the target population is a common problem in surveys. Undercoverage occurs when some population units are not included in the sampling frame and have no chance of being selected into the sample. Almost all surveys are subject to some amount of undercoverage, and NAAL and SAAL are no exception. A calibration adjustment to the weights accounted for any undercoverage and balanced the samples within each SAAL state prior to the compositing process. For this step, the entire sample was divided into the NAAL and SAAL sample in the six SAAL states. The NAAL sample in the remaining 44 states and the District of Columbia was excluded from this step because there was no SAAL sample in those states. After compositing, the combined NAAL and SAAL household sample weights were calibrated through a raking adjustment process (refer to section 12.1.6.2).

The creation of the control totals used for the calibration adjustment is discussed in section 12.1.4.1. The calibration adjustments are discussed in section 12.1.4.2.

12.1.4.1 Control Totals

Control totals were computed for the purpose of calibrating the sample weights within the six SAAL states prior to compositing. The totals were computed from the 2003 CPS March Supplement. For each sample, control totals were computed for the following variables: MSA status, age, gender, education, country of birth, race/ethnicity, and national certainty status of the PSU. The number of variables to use was limited because the external source of the control totals needed to have the exact same wording of questions as the NAAL and no missing NAAL responses. Furthermore, not all of these variables were used for each state sample because of small sample sizes in certain domains. However, the variables used for the calibration step were defined to the finest classification that the data allow. Also the effectiveness of calibration methods (raking in particular) depends on the relationship between the auxiliary variables used in calibration and the survey estimates (Brick et al. 2003). Table 12-6 displays the resulting variables involved in the calibration process.

12.1.4.2 Calibration

Calibration is commonly used in sample surveys to reduce the mean square error of estimates and to create consistency with statistics from other studies. However, the primary reason for calibration in the setting of NAAL is to provide a common base for the NAAL and SAAL samples in each of the six SAAL states before applying the composite weighting factors. The trimmed background questionnaire weights for the six states were calibrated to the 2003 CPS March Supplement control totals. Respondents who completed the background questionnaire were included in the calibration. Literacy-related nonrespondents were also included because they are part of the target population from which the control totals were derived. Variables critical to the weighting were recoded and imputed, as necessary, before the calculation of base weights.

Table 12-6. Variables involved in the calibration process prior to compositing, by sample: 2003

Sample[1]	MSA status	Age	Gender[2]	Education[3]	Country of birth	Race/ethnicity	NAAL certainty PSU
NAAL Kentucky		16–49, 50+	M, F	<HS, HS, +		Hispanic/Black, other	
NAAL Maryland		16–49, 50+	M, F	<HS/HS, +		Hispanic/Black, other	
NAAL Massachusetts		16–49, 50+	M, F	<HS/HS, +		Hispanic/Black, other	
NAAL Missouri		16–49, 50+	M, F	<HS/HS, +			
NAAL New York						Hispanic/Black, other	Certainty, noncertainty
NAAL Oklahoma		16–49, 50+	M, F	<HS, HS, +			
SAAL Kentucky	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +		Hispanic/Black, other	
SAAL Maryland		16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	
SAAL Massachusetts		16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	
SAAL Missouri	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +		Hispanic/Black, other	
SAAL New York						Hispanic/Black, other	
SAAL Oklahoma	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +		Hispanic/Black, other	

[1] NAAL XX means the national NAAL sample cases in state XX.
[2] M: males; F: females.
[3] <HS: less than high school; HS: high school diploma or equivalent; +: more than high school; <HS/HS: either less than high school or high school diploma or equivalent.
NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

A raking procedure (i.e., iterative poststratification) was used for the calibration. In raking, categories are formed from certain variables and the weights are calibrated to control totals for each category. In some instances, such cross-tabulations may contain sparse cells, or population distributions may be known for the marginal but not the joint distributions for variables used to define the weighting classes. Typically, raking is conducted when the control totals for interior cells of a cross-tabulation are unknown or sample sizes in some cells are too small for efficient estimation. Raking is related to poststratification in that it poststratifies (or calibrates) to marginal population totals of several variables (or raking dimensions) in an iterative manner. Oh and Scheuren (1987) provide a concise description of the raking procedure and its properties.

A raked weight was calculated for each respondent as follows. Let N_γ denote the population count in the raking dimension category γ as obtained from the 2003 CPS March Supplement, as discussed in section 12.1.4.1. Let $\hat{N}_\gamma$ be the corresponding survey estimate obtained by using the survey weights prior to raking (as calculated below):

$$\hat{N}_\gamma = \sum_{i=SPL(\gamma)} W_{ijkl}^{base,BQ} F_\beta^{nr,BQ} F_{ijkl}^{trim,BQ},$$

where

$W_{ijkl}^{base,BQ} F_\beta^{nr,BQ} F_{ijkl}^{trim,BQ}$ = the sample weight for person l, reflecting all weighting adjustments prior to raking, and

SPL(γ) = the set of background questionnaire respondents and literacy-related background questionnaire nonrespondents in raking dimension category γ.

The adjustment factor for raking dimension category γ is given by $F_\gamma = N_\gamma / \hat{N}_\gamma$. The same process is applied for each raking dimension, each time using the adjusted weights from the previous dimension. This is done iteratively until the sums of the adjusted weights equal all control totals. The raking processes all converged in less than 15 iterations.

For simplicity, the raking factor can be denoted as $F_\gamma^{Cal,BQ}$, where γ can denote each of the interior cells defined by the raking dimensions shown in table 12-6.

At this stage of the weighting process, the calibration is done only for cases in the SAAL states to provide a common base for the NAAL and SAAL samples, prior to compositing the weights. Therefore, the calibration factor was set equal to 1 for all persons outside the six SAAL states. The calibration factor is then applied to the sample weights to create the weights used in the composite weighting process:

$$W_{ijkl}^{cal,BQ} = W_{ijkl}^{base,BQ} F_{\beta}^{nr,BQ} F_{ijkl}^{trim,BQ} F_{\gamma}^{Cal,BQ}$$

12.1.5 Compositing Data from the National and State Household Components

The original plan for the 1992 National Adult Literacy Survey (NALS) was to consider the national and state samples as two separate surveys so that national statistics would be prepared from the national sample only and state data would be prepared from the state samples only. An evaluation of the 1992 NALS data showed that the increased sample size resulting from the combination of the two samples improved precision for both state and national estimates (Burke et al. 1994). The combined sample had the additional advantage of producing a single database for state and national statistics. Therefore, the NAAL and SAAL samples were combined for the 2003 NAAL as well. The method of combining data from the state and national samples is referred to as composite estimation.

The standard theoretical foundation of composite estimation requires a knowledge of variances of the statistics of interest, in this case, the literacy scores. This information is necessary to produce the parameters used to combine data from various surveys in a way that minimizes the variances of the composite estimates. After the literacy data became available from the 1992 NALS, new compositing factors were computed for a selected set of statistics. (Refer to section 11.2.4 of the Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey [NCES 2001-457] [Kirsch et al. 2001]). Also, at that time an approach was developed for creating efficient compositing factors for the next national adult literacy study.

Section 12.1.5.1 describes the composite estimation procedure used for the 2003 NAAL. The calculation of the compositing factors is discussed in section 12.1.5.2.

12.1.5.1 Composite Estimation Procedure

In general, the composite estimator for a combined state sample is given by

$$\hat{Y} = \beta \hat{Y}_{st} + (1 - \beta) \hat{Y}_{nt},$$

where

$\hat{Y}$ = the composite estimate for variable Y;

β = the composite factor (0 < < 1);

$\hat{Y}_{st}$ = the estimate of Y coming from the state sample; and

$\hat{Y}_{nt}$ = the estimate of Y coming from the national sample.

The variance of a composite estimator will be smaller than the variance of both the national and state estimates if appropriate composite factors are used. Optimal factors can be found when unbiased estimators exist for the two components and approximate estimates of their variances are available. It should be noted that a composite estimator produces unbiased estimates for any value of ˜ The optimum value of . is the one that results in the lowest variance.

As stated above, the national and state samples were selected independently and each could, thus, produce unbiased estimates of subdomain statistics for persons 16 years and older. Therefore, factors could be derived to produce composite estimators with variances that were smaller than those of either of the two estimates. For statistic Y, the optimal composite factor for a particular state is

$$\beta = \frac{V(\hat{Y}_{nt})}{V(\hat{Y}_{nt}) + V(\hat{Y}_{st})}, \tag{1}$$

where

$V(\hat{Y}_{nt})$ = the variance of the estimate of Y coming from the national sample and

$V(\hat{Y}_{st})$ = the variance of the estimate of Y coming from the state sample.

A different optimal value of might be found for each statistic of interest. However, data analyses would be complicated if item-specific values of were used because items would not add up to totals, or totals derived by summing different items would not agree. Consequently, the goal for NAAL was to associate with each sample person a single compositing factor that although not precisely optimal

for any particular statistic would be robust enough to enhance the precision of virtually all composited statistics. This objective was accomplished by focusing on aspects of the sample design that were likely to affect the variance, regardless of the choice of statistic.

12.1.5.2 Estimating the Compositing Factors

Two aspects of the design should be reflected in the compositing factors. One is the distinction between cases coming from national certainty or noncertainty PSUs. The next design aspect is the oversampling of Blacks and Hispanics in the national sample. The oversampling introduced variability in the weights and increased the design effect for cases coming from the national sample. To best reflect these design features, separate compositing factors (denoted by δ) were created from the combinations of state, certainty status of national PSUs, and race/ethnicity.

The compositing factor in equation (1) can be rewritten as follows:

$$\overline{\beta}_{\delta} = \frac{Ratio_{\delta}(\text{var})}{Ratio_{\delta}(\text{var})+1},$$

where $Ratio_{\delta}$ (var) is the ratio of the variances from subgroup δ coming from the state and national samples. This ratio is calculated differently for PSUs that are certainties in the national sample and those that are not certainties in the national sample:

$$Ratio_{\delta}(\text{var}) = \overline{R}_{qg}\frac{n_{(st)\delta}}{n_{(nt)\delta}} \quad \text{for national certainty PSU and}$$

$$Ratio_{\delta}(\text{var}) = \overline{R}_{qg} \times \left\{ \frac{\dfrac{\overline{P}_{(nt)g}}{m_{(nt)\delta}} + \dfrac{1-\overline{P}_{(nt)g}}{n_{(nt)\delta}}}{\dfrac{\overline{P}_{(st)g}}{m_{(st)\delta}} + \dfrac{1-\overline{P}_{(st)g}}{n_{(st)\delta}}} \right\} \quad \text{otherwise,}$$

where

n(nt)δ = *the number of respondents in subgroup δ of the national sample;*

n(st)δ = *the number of respondents in subgroup δ of the state sample;*

m(nt)δ = *the number of PSUs in subgroup δ of the national sample;*

$m(st)\delta$ = *the number of PSUs in subgroup δ of the state sample;*

$\overline{R}_{qg}$ = *the average value of the ratio of the unit variances for sample cases in race/ethnicity category g in national PSUs with certainty status q;*

$\overline{P}_{(nt)g}$ = *the average proportion of the national unit variance for subgroup g coming from the between-PSU component; and*

$\overline{P}_{(st)g}$ = *the average proportion of the state unit variance for subgroup g coming from the between-PSU component.*

The values of $\overline{R}_{qg}$, $\overline{P}_{(nt)g}$, and $\overline{P}_{(st)g}$ are parameters computed from the postweighting 1992 NALS analysis. These values, along with the calculations of Ratioδ (var) and $\overline{\beta}_{\delta}$, are shown in table 12-7.

Table 12-7. Calculations of Ratioδ (var) and $\overline{\beta}_{\delta}$, by certainty in national sample, race/ethnicity, and state: 2003

Certainty in national sample	Race/ethnicity	State	n(nt)δ	m(nt)δ	$\overline{R}_{qg}$	$\overline{P}_{(nt)g}$	n(st)δ	m(st)δ	$\overline{P}_{(st)g}$	Ratioδ(var)	$\overline{\beta}_{\delta}$	$_{1-}\overline{\beta}_{\delta}$
No	Hispanic	KY	3	2	1.1322	0.0032	26	18	0.0004	9.8264	0.9076	0.0924
No	Black	KY	53	2	0.7365	0.0054	113	18	0.0041	1.7487	0.6362	0.3638
No	Other	KY	138	2	0.8115	0.0008	1,212	18	0.0006	7.2272	0.8785	0.1215
No	Hispanic	MD	2	1	1.1322	0.0032	38	13	0.0004	21.5641	0.9557	0.0443
No	Black	MD	75	1	0.7365	0.0054	247	13	0.0041	3.1615	0.7597	0.2403
No	Other	MD	53	1	0.8115	0.0008	601	13	0.0006	9.3317	0.9032	0.0968
Yes	Hispanic	MA	18	1	1.1275	—	28	1	—	1.7539	0.6369	0.3631
Yes	Black	MA	6	1	0.7363	—	21	1	—	2.5771	0.7204	0.2796
Yes	Other	MA	81	1	0.8118	—	331	1	—	3.317	0.7684	0.2316
No	Hispanic	MA	33	1	1.1322	0.0032	43	5	0.0004	1.621	0.6185	0.3815
No	Black	MA	17	1	0.7365	0.0054	29	5	0.0041	1.3386	0.5724	0.4276
No	Other	MA	89	1	0.8115	0.0008	378	5	0.0006	3.5312	0.7793	0.2207
No	Hispanic	MO	1	1	1.1322	0.0032	25	12	0.0004	28.2927	0.9659	0.0341
No	Black	MO	0	0	0.7365	0.0054	99	12	0.0041	†	1.0000	—
No	Other	MO	84	1	0.8115	0.0008	800	12	0.0006	7.9293	0.8880	0.1120
Yes	Hispanic	NY	181	4	1.1275	—	90	4	—	0.5606	0.3592	0.6408
Yes	Black	NY	187	4	0.7363	—	75	4	—	0.2953	0.2280	0.7720
Yes	Other	NY	150	4	0.8118	—	203	4	—	1.0986	0.5235	0.4765
No	Hispanic	NY	27	4	1.1322	0.0032	15	8	0.0004	0.6403	0.3904	0.6096
No	Black	NY	92	4	0.7365	0.0054	13	8	0.0041	0.1161	0.1041	0.8959
No	Other	NY	386	4	0.8115	0.0008	311	8	0.0006	0.6881	0.4076	0.5924
No	Hispanic	OK	8	2	1.1322	0.0032	60	12	0.0004	8.5593	0.8954	0.1046
No	Black	OK	5	2	0.7365	0.0054	70	12	0.0041	10.1925	0.9107	0.0893
No	Other	OK	184	2	0.8115	0.0008	960	12	0.0006	4.3366	0.8126	0.1874

— Not available.
† Not applicable.
NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

12.1.6 Computing Final Weights

The final weights were created by applying the composite factors to the calibrated weights (section 12.1.6.1) and then raking the weights to control totals. The raking process included the following sequence of subtasks: raking, trimming, and reraking (refer to section 12.1.6.2).

12.1.6.1 Compositing the NAAL and SAAL Samples

After calculating the compositing factor, $\overline{\beta}_{\delta}$, the composited weight, $W_{ijkl}^{comp,BQ}$, was computed as follows:

$W_{ijkl}^{comp,BQ}$ = $\overline{\beta}_{\delta} W_{ijkl}^{cal,BQ}$ for person l in subgroup δ, and associated with the SAAL sample;

= $(1-\overline{\beta}_{\delta}) W_{ijkl}^{cal,BQ}$ for person l in subgroup δ, and associated with the national NAAL household sample in SAAL states; and

= $W_{ijkl}^{cal,BQ}$ for person l in non-SAAL states.

12.1.6.2 Raking Composite Weights

The final step in weighting was to rake the composited weights to control totals. The raking process was completed for the entire sample in a manner similar to the calibration performed before compositing for the six SAAL states. The process included the following steps: creating control totals, raking, trimming, and reraking. The creation of the control totals for the calibration adjustment is discussed in section 12.1.6.2.1. The weighting steps of raking, trimming, and reraking are discussed in section 12.1.6.2.2.

12.1.6.2.1 Control Totals

Control totals were computed for the purpose of creating the final weights for the entire combined NAAL and SAAL household sample. The totals were computed from the 2003 CPS March Supplement for each SAAL state and the remainder of the United States. For each sample, control totals were computed for the following variables: MSA status, age, gender, education, country of birth, and race/ethnicity. Census region was used for the sample containing the remainder of the United States. Not

all of these variables were used for each state sample because of small sample sizes in certain domains. However, the variables used for the raking step were defined to the finest classification that the data allow. Further discussion as to the selection of raking variables is in section 12.1.4.1. Table 12-8 displays the variables involved in the raking process.

12.1.6.2.2 Final Adjustments

The final steps in the weighting process were raking, trimming, and raking again to recalibrate the weights to the control totals. A general overview of trimming is provided in section 12.1.3.3, and a general overview of raking is provided in section 12.1.4.2.

Respondents who completed the background questionnaire were included in the raking process. Literacy-related nonrespondents were also included because (1) the reasons for nonparticipation have been found to be highly related to literacy results (NAAL 2001) and (2) they are part of the target population from which the control totals were derived. Table 12-9 summarizes the raking factors for the first round of raking. In general, the mean raking factors are near 1.0 for SAAL states and 1.13 for the rest of the nation. By domain, the means are more variable as one would expect because the sample sizes are smaller for domains in general. The table also shows the range of the raking factors by state and by raking dimension. Convergence in raking is generally achieved within the first few iterations. A maximum number of 15 iterations was preset for the reason that any further processing would be for naught because convergence would be unlikely. All raking processes converged in less than 15 iterations.

After raking, the trimming process was repeated to adjust extreme weights created after raking. In this step, fewer than 1 percent of the weights were reduced.

The last step was a second round of raking. Table 12-10 summarizes the raking factors for the second round of raking after the compositing procedure was applied. As shown in the table, this raking step had little effect on the weights as the adjustment factors are near 1.0.

Table 12-8. Variables involved in the raking process after compositing, by sample: 2003

Sample	MSA status	Age	Gender[1]	Education[2]	Country of birth	Race/ethnicity	Census region
Kentucky	MSA, non-MSA	16–29, 30–49, 50+	M, F	<HS, HS, +		Hispanic/Black, other	
Maryland		16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	
Massachusetts		16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	
Missouri	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +		Hispanic/Black, other	
New York	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	
Oklahoma	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +		Hispanic/Black, other	
Remainder of U.S.	MSA, non-MSA	16–29, 30–49, 50–69, 70+	M, F	<HS, HS, +	U.S., other	Hispanic/Black, other	Northeast, Midwest, South, West

[1] M: Males, F: Females.
[2] <HS: less than high school; HS: high school diploma or equivalent; +: more than high school.
NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Table 12-9. Raking factors for the first round of raking after compositing, by state and domain: 2003

Domain	Kentucky				Maryland				Massachusetts				Missouri			
	n	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean
Overall	1,545	0.90	1.05	0.99	1,016	0.85	1.24	0.99	1,074	0.76	1.20	0.99	1,009	0.66	1.19	0.96
Region																
Northeast	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
Midwest	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
South	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
West	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
MSA status																
Non-MSA	807	0.93	1.05	1.01	†	†	†	†	†	†	†	†	400	0.66	0.79	0.75
MSA	738	0.90	1.02	0.98	†	†	†	†	†	†	†	†	609	0.99	1.19	1.10
Age																
16–29	361	0.95	1.05	1.01	222	0.95	1.24	1.06	224	0.82	1.20	1.04	233	0.66	1.14	0.95
30–49	591	0.90	1.00	0.98	406	0.85	1.11	0.95	465	0.76	1.11	0.97	385	0.67	1.16	0.97
50–69	593	0.93	1.03	1.00	281	0.89	1.15	0.99	257	0.78	1.12	0.98	267	0.67	1.16	0.96
70+					107	0.90	1.18	1.00	128	0.83	1.19	1.04	124	0.76	1.19	0.98
Gender																
Male	637	0.90	1.05	0.99	413	0.87	1.24	1.00	486	0.78	1.20	1.00	458	0.68	1.19	0.96
Female	908	0.91	1.05	1.00	603	0.85	1.21	0.98	588	0.76	1.19	0.99	551	0.66	1.18	0.96
Education																
Less than HS[1]	363	0.90	1.02	0.98	151	0.85	1.16	0.95	176	0.76	1.11	0.93	213	0.69	1.19	0.95
HS	518	0.93	1.04	1.00	256	0.91	1.20	0.99	209	0.83	1.20	1.02	248	0.68	1.17	0.96
More than HS	664	0.94	1.05	1.00	609	0.90	1.24	1.00	689	0.82	1.19	1.00	548	0.66	1.15	0.97
Born in U.S.																
No	†	†	†	†	130	0.97	1.24	1.10	244	0.76	1.00	0.87	†	†	†	†
Yes	†	†	†	†	886	0.85	1.09	0.97	830	0.92	1.20	1.03	†	†	†	†
Race/ethnicity																
Hispanic	195	0.904	1.01	0.95	362	0.85	1.20	0.97	195	0.84	1.20	0.98	125	0.66	1.09	0.99
Non-Hispanic Other	1,350	0.94	1.05	1.00	654	0.88	1.24	1.00	879	0.76	1.10	1.00	884	0.72	1.19	0.96

See notes at end of table.

Table 12-9. Raking factors for the first round of raking after compositing, by state and domain: 2003—Continued

Domain	New York				Oklahoma				Remaining sample			
	n	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean
Overall	1,730	0.61	4.30	1.03	1,287	0.95	1.09	0.99	10,880	0.65	1.77	1.13
Region												
Northeast	†	†	†	†	†	†	†	†	961	0.86	1.77	1.25
Midwest	†	†	†	†	†	†	†	†	2,603	0.65	1.41	1.00
South	†	†	†	†	†	†	†	†	4,422	0.75	1.65	1.14
West	†	†	†	†	†	†	†	†	2,894	0.80	1.76	1.19
MSA status												
Non-MSA	73	1.84	4.30	2.73	515	0.97	1.09	1.00	2,043	0.65	1.40	0.98
MSA	1,657	0.61	1.90	0.95	772	0.95	1.08	0.99	8,837	0.81	1.77	1.16
Age												
16–29	424	0.64	3.88	0.99	355	0.99	1.09	1.01	2,893	0.67	1.67	1.14
30–49	669	0.65	3.92	1.04	429	0.95	1.06	0.98	4,316	0.66	1.61	1.10
50–69	437	0.61	3.70	1.00	337	0.97	1.08	1.00	2,573	0.65	1.63	1.12
70+	200	0.71	4.30	1.12	166	0.97	1.08	0.99	1,098	0.80	1.77	1.23
Gender												
Male	706	0.70	4.30	1.10	557	0.96	1.09	1.00	4,771	0.68	1.77	1.14
Female	1,024	0.61	3.83	0.97	730	0.95	1.09	0.99	6,109	0.65	1.74	1.12
Education												
Less than HS[1]	441	0.61	2.44	0.83	314	0.95	1.09	1.00	2,928	0.69	1.37	1.06
HS	389	1.08	4.30	1.44	355	0.96	1.08	1.00	2,734	0.86	1.77	1.36
More than HS	900	0.75	3.00	0.94	618	0.96	1.09	0.99	5,218	0.65	1.33	1.05
Born in U.S.												
No	470	0.65	3.83	0.96	†	†	†	†	2,247	0.65	1.72	1.08
Yes	1,260	0.61	4.30	1.05	†	†	†	†	8,633	0.66	1.77	1.14
Race/ethnicity												
Hispanic	313	0.64	3.83	0.91	143	1.03	1.09	1.06	2,596	0.65	1.57	1.08
Non-Hispanic Black Only	367	0.76	1.90	1.10	†	†	†	†	2,402	0.72	1.67	1.15
Other	1,050	0.61	4.30	1.03	1,144	0.95	1.02	0.99	5,882	0.72	1.77	1.14

† Not applicable.
[1] High school.
NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Table 12-10. Raking factors for the second round of raking after compositing, by state and domain: 2003

Domain	Kentucky				Maryland				Massachusetts				Missouri			
	n	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean
Overall	1,545	1.00	1.02	1.00	1,016	0.99	1.01	1.00	1,074	0.98	1.07	1.01	1,009	0.98	1.04	1.01
Region																
Northeast	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
Midwest	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
South	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
West	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
MSA status																
Non-MSA	807	1.00	1.02	1.00	†	†	†	†	†	†	†	†	400	0.98	1.03	1.00
MSA	738	1.00	1.02	1.00	†	†	†	†	†	†	†	†	609	1.00	1.04	1.01
Age																
16–29	361	1.00	1.02	1.00	222	0.99	1.01	1.00	224	0.99	1.07	1.02	233	0.99	1.04	1.01
30–49	591	1.00	1.02	1.00	406	0.99	1.01	1.01	465	0.98	1.06	1.01	385	0.99	1.04	1.00
50–69	593	1.00	1.01	1.01	281	0.99	1.01	1.00	257	0.98	1.06	1.00	267	0.98	1.03	1.00
70+	†	†	†	†	107	0.99	1.01	1.00	128	0.98	1.05	1.00	124	0.98	1.03	1.00
Gender																
Male	637	1.00	1.02	1.00	413	0.99	1.01	1.00	486	0.98	1.07	1.01	458	0.98	1.04	1.00
Female	908	1.00	1.02	1.00	603	0.99	1.01	1.00	588	0.98	1.06	1.01	551	0.99	1.04	1.01
Education																
Less than HS	363	1.00	1.02	1.00	151	0.99	1.01	1.00	176	0.98	1.04	1.01	213	0.98	1.01	1.00
HS	518	1.00	1.01	1.00	256	0.99	1.01	1.00	209	1.00	1.07	1.02	248	1.01	1.04	1.03
More than HS	664	1.00	1.02	1.00	609	0.99	1.01	1.00	689	0.98	1.05	1.00	548	0.98	1.01	1.00
Born in U.S.																
No	†	†	†	†	130	0.99	1.01	1.00	244	0.98	1.06	1.01	†	†	†	†
Yes	†	†	†	†	886	0.99	1.01	1.00	830	0.99	1.07	1.01	†	†	†	†
Race/ethnicity																
Hispanic	195	1.01	1.02	1.01	362	1.00	1.01	1.01	195	1.02	1.07	1.04	125	0.98	1.04	1.01
Non-Hispanic	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
Other	1,350	1.00	1.01	1.00	654	0.99	1.01	1.00	879	0.98	1.03	1.00	884	0.98	1.04	1.00

See notes at end of table.

Table 12-10. Raking factors for the second round of raking after compositing, by state and domain: 2003—Continued

Domain	New York				Oklahoma				Remaining sample			
	n	Min	Max	Mean	*n*	Min	Max	Mean	*n*	Min	Max	Mean
Overall	1,730	0.97	1.12	1.02	1,287	1.00	1.00	1.00	10,880	0.994	1.01	1.00
Region												
Northeast	†	†	†	†	†	†	†	†	961	0.996	1.01	1.00
Midwest	†	†	†	†	†	†	†	†	2,603	0.994	1.01	1.00
South	†	†	†	†	†	†	†	†	4,422	0.997	1.01	1.00
West	†	†	†	†	†	†	†	†	2,894	0.994	1.01	1.00
MSA status												
Non-MSA	73	1.00	1.12	1.05	515	1.00	1.00	1.00	2,043	0.994	1.01	1.00
MSA	1,657	0.97	1.12	1.02	772	1.00	1.00	1.00	8,837	0.996	1.01	1.00
Age												
16–29	424	0.98	1.12	1.02	355	1.00	1.00	1.00	2,893	0.995	1.01	1.00
30–49	669	0.97	1.11	1.02	429	1.00	1.00	1.00	4,316	0.994	1.00	1.00
50–69	437	0.99	1.12	1.03	337	1.00	1.00	1.00	2,573	0.995	1.01	1.00
70+	200	0.98	1.11	1.02	166	1.00	1.00	1.00	1,098	0.994	1.01	1.00
Gender												
Male	706	0.99	1.12	1.04	557	1.00	1.00	1.00	4,771	0.995	1.01	1.00
Female	1,024	0.97	1.11	1.01	730	1.00	1.00	1.00	6,109	0.994	1.01	1.00
Education												
Less than HS	441	0.97	1.03	1.00	314	1.00	1.00	1.00	2,928	0.994	1.01	1.00
HS	389	1.06	1.12	1.08	355	1.00	1.00	1.00	2,734	0.999	1.01	1.00
More than HS	900	0.99	1.05	1.01	618	1.00	1.00	1.00	5,218	0.994	1.01	1.00
Born in U.S.												
No	470	0.99	1.12	1.03	†	†	†	†	2,247	0.996	1.01	1.00
Yes	1,260	0.97	1.12	1.02	†	†	†	†	8,633	0.994	1.01	1.00
Race/ethnicity												
Hispanic	313	0.97	1.12	1.02	143	1.00	1.00	1.00	2,596	0.994	1.01	1.00
Non-Hispanic	367	0.97	1.12	1.02	†	†	†	†	2,402	0.994	1.01	1.00
Other	1,050	0.97	1.12	1.02	1,144	1.00	1.00	1.00	5,882	0.995	1.01	1.00

† Not applicable.

NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.

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

The final weights were computed as the product of the composited background questionnaire weight and the raking and trimming factors, as shown below:

$$W_{ijkl}^{final,BQ} = W_{ijkl}^{comp,BQ} F_{\gamma}^{rake1,BQ} F_{ijkl}^{trim2,BQ} F_{\gamma}^{rake2,BQ},$$

where

$F_{\gamma}^{rake1,BQ}$ = the first-round raking factor for sample persons in raking cell γ;

$F_{ijkl}^{trim2,BQ}$ = the trimming factor after compositing for sample person l; and

$F_{\gamma}^{rake2,BQ}$ = the second-round raking factor for sample persons in raking cell γ.

The distribution of the final background questionnaire weights is shown in table 12-11 for the combined NAAL and SAAL sample. The coefficient of variation, shown in the table, measures the variation of the sampling weights, which can affect the precision of survey outcome statistics. The total combined NAAL and SAAL sample has a relatively high coefficient of variation of the weights due to a mixture of oversampling Blacks and Hispanics in the national sample and higher sampling rates for SAAL states.

Table 12-11. Distribution of final background questionnaire weights, by domain: 2003

Domain	Sample size	Minimum	Sum	Median	Maximum	Coefficient of variation
Total	18,541	149	221,021,328	7,697	72,267	95
Kentucky	1,545	149	3,222,654	2,052	6,482	44
Maryland	1,016	887	4,228,643	3,856	11,899	50
Massachusetts	1,074	568	5,155,801	4,316	14,268	53
Missouri	1,009	739	4,355,187	3,835	11,673	60
New York	1,730	809	15,119,508	7,008	27,174	67
Oklahoma	1,287	221	2,706,561	1,959	5,963	46
Rest of U.S.	10,880	1,567	186,232,851	13,176	72,267	69

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

12.1.6.3 Quality Control

When several stages of adjustments are used to produce sampling weights, it is essential that quality control (QC) checks are done throughout the weighting process. The checks done for NAAL

included reviewing the computer code used to calculate the weights, validating the implementation of weighting specifications, and calculating and checking descriptive statistics on the weights.

A standardized weighting system was used to perform the tasks of nonresponse adjustment, raking, and the creation of replicate weights for variance estimation. The system has been used on numerous surveys and has been well tested through careful review procedures.

Despite careful review of the specifications and program, there still needs to be careful review of the resulting weights. After each weighting step, the weighted totals and percent distribution for several key domains was compared to the weighted totals and percent distribution prior to the weighting step. This allowed the statistician to identify any large changes in distribution and investigate into the reasons for the changes.

12.1.7 Replicate Weights

Variance estimation must take into account the sample design. In particular, the sampling variance estimate for any statistic should account for the effects of clustering; the use of nonresponse, trimming, and poststratification adjustments; and the component of sampling variability arising from the variation in the weights used to compute the statistic. Treating the data as a simple random sample will produce underestimates of the true sampling variability for the NAAL area sample design.

The stratified jackknife method can be used to estimate the variance for most statistics. Jackknifing estimates the sampling variability of any statistic Y, as the sum of components of variability that may be attributed to individual pairs of first-stage sampling units. The variance attributed to a particular pair is measured by estimating how much the value of the statistic would change if only one unit in the pair had been sampled. When using replication techniques such as jackknifing to calculate standard errors, it is necessary to establish a number of subsamples (or replicates) from the full sample, calculate the estimate from each subsample, and sum the squared difference of each replicated estimate from the full-sample estimate. The 61 replicates formed for the NAAL and SAAL combined household sample provided the degrees of freedom necessary for the production of stable estimates of variance.

Three steps were involved in facilitating variance estimation: (1) forming the replicates, (2) constructing the replicate weights, and (3) computing estimates of variance for survey statistics. Under the stratified jackknife approach employed for NAAL, the number of replicates is equal to the number of

degrees of freedom, which is directly related to the stability of variance estimates (Rust 1985). The formation of replicates is discussed in detail in sections 12.1.7.1 through 12.1.7.3. After the replicates were formed, a replicate factor was constructed for each variance stratum. Let $f_{ijk}(r)$ denote the rth replicate factor for sampled dwelling unit k in variance unit i′ of variance stratum h′ (the prime symbol is used to distinguish between the variance unit and PSU i, and also between the variance stratum and PSU stratum h). Then, in general,

$$f_{ijk}(r) = \begin{cases} 2 & \text{if } h' = r \text{ and } i' = 1; \\ 0 & \text{if } h' = r \text{ and } i' = 2; \text{ and} \\ 1 & \text{if } h' \neq r, \end{cases}$$

and the replicated screener base weight, $W_{ijk}^{base,SCR}(r)$, was obtained as

$$W_{ijk}^{base,SCR}(r) = W_{ijk}^{base,SCR} f_{ijk}(r)$$

for r = 1, 2, ~, 61. For SAAL, 13 to 19 replicate weights were formed, depending on the state.

After obtaining a screener base weight for each replicate, all remaining full-sample weighting steps leading to the final person weight were performed on each replicate. The repetition of the various weight adjustment procedures on each set of replicate base weights ensures that the impact of these procedures on the sampling variance of the estimator Y is appropriately reflected in the variance estimator, v(Y).

12.1.7.1 Replicate Weights for the National Sample

The national sample contained 100 PSUs, 16 of which were selected with certainty. The remaining 84 PSUs were selected one per stratum with probabilities proportional to size. Replicates were formed by pairing first-stage sampled units; that is, segments were paired in PSUs selected with certainty, and whole PSUs were paired in noncertainty strata. For the 100-PSU sample, the natural pairing led to 61 replicates.

12.1.7.2 Replicate Weights for the State Sample

An independent sample of PSUs was selected in each of the six participating states. The largest PSUs were taken with certainty. Within each state, the remaining PSUs were grouped into strata, and from each stratum a single PSU was sampled with probability proportional to size. In PSUs selected with certainty, segments were paired to form replicates. The number of replicates formed for each state was as follows:

Kentucky	16
Maryland	15
Massachusetts	13
Missouri	13
New York	19
Oklahoma	13

For Maryland and Massachusetts, segments were the first-stage sampling units. Therefore, the selected segments were grouped and paired to form replicates. In addition, Kentucky and Oklahoma each had one triplet; that is, one variance stratum had three variance units. The triplets generate one additional replicate, while affecting two replicates in total, using a factor of 1.5 to construct the replicate weights.

12.1.7.3 Final Replication for the National and State Samples

The NAAL analysis combined data from a nationally representative sample of 100 PSUs with data from 6 independently selected state PSU samples. The threefold objective of the replication scheme was to (1) reflect the actual sample design of each sample; (2) ensure the production of stable estimates of standard errors by having sufficient degrees of freedom for national estimates, individual SAAL state estimates, and regional estimates; and (3) limit the total number of replicates so that variance estimation could be done more efficiently. The general approach in setting up the replication was to devise an appropriate scheme for each component of the sample, the national sample, and the six states, and then to collapse replicates to a reasonable number.

12.1.7.3.1 Active Replicates

A total of 150 replicates had been formed at this point: 61 from the national sample and 89 from the state samples. These replicates reflected the actual design of each sample and provided sufficient degrees of freedom to produce stable estimates of variance for the nation, each state, and the four census regions. However, using 150 replicates to estimate variances would greatly increase the computer processing time for data analysis, while providing only a slight gain in the precision of the overall variance estimates. This is analogous to increases in sample size providing diminishing returns with respect to the precision of estimates. Refer to Rust and Rao (1996) for further discussion. Therefore, the number of replicates was collapsed to 61. To preserve the total number of replicates for each state, replicates from the same state were never collapsed. To the extent possible, replicates from the same region were not combined either.

Table 12-12 presents the results of the replication scheme, showing which replicates are active for the major subdomains of analysis.

12.1.7.3.2 Replicated Control Totals

As mentioned above, the 2003 CPS March Supplement was used to create the population control totals for the household component. In general, control totals derived from the CPS have some variance associated with them because the CPS is a survey (not a census). Usually, the sampling errors are ignored when using the control totals for the U.S. population (or main subgroups of the U.S. population) because the very large size of the CPS sample results in very small variances. However, the state sample sizes in the CPS are smaller, and the variances associated with the state-level control totals are relatively high. This section describes the approach used to add variation to the replicate totals, reflecting the CPS variances. The CPS variances are measured using generalized variance functions (GVFs). The GVF model is a regression model fit to the survey relative variance, V2, as follows,

$$V^2=A+B/X$$

where A is the intercept, B is the slope, and X is a set of survey estimates. The resulting parameters, A and B, can be used to approximate the standard error associated with any survey estimate X. The GVF parameters are found in the source and accuracy statement of the 2003 CPS March Supplement.

Table 12-12. Active replicates for subdomains of the NAAL analysis file, by replicate, region, and state: 2003

Replicate	Region and state										
	U.S.	Northeast	Midwest	South	West	Massachusetts	New York	Kentucky	Oklahoma	Maryland	Missouri
1	x	x		x				x			
2	x	x		x		x		x			
3	x	x		x			x	x			
4	x	x		x			x	x			
5	x	x		x			x	x			
6	x	x	x	x			x	x			
7	x	x	x	x			x	x			
8	x	x	x	x			x	x			
9	x	x	x	x			x	x			
10	x	x	x	x			x			x	
11	x	x	x	x			x			x	
12	x	x	x	x			x			x	
13	x	x	x	x			x			x	
14	x	x		x			x				
15	x	x		x			x				
16	x	x		x			x				
17	x	x		x			x				
18	x	x		x			x			x	
19	x	x		x			x				
20	x	x		x			x				
21	x	x		x			x				
22	x	x		x			x	x			
23	x	x		x			x				
24	x	x		x			x		x		
25	x	x		x		x					
26	x	x		x	x	x				x	
27	x	x		x	x	x				x	
28	x	x		x	x	x				x	
29	x	x		x	x	x				x	
30	x	x		x	x	x				x	
31	x	x		x	x	x				x	
32	x	x		x						x	

See notes at end of table.

Table 12-12. Active replicates for subdomains of the NAAL analysis file, by replicate, region, and state: 2003—Continued

Replicate	Region and state										
	U.S.	Northeast	Midwest	South	West	Massachusetts	New York	Kentucky	Oklahoma	Maryland	Missouri
33	x	x	x	x		x				x	
34	x	x	x	x		x				x	x
35	x	x	x	x		x				x	
36	x	x		x		x					
37	x	x		x		x					
38	x	x		x		x					
39	x		x	x				x			x
40	x		x	x					x		x
41	x		x	x	x					x	x
42	x		x	x	x				x		x
43	x	x	x	x		x			x		x
44	x	x	x	x					x		x
45	x	x	x	x			x		x		x
46	x	x	x	x			x		x		x
47	x	x	x	x			x		x		x
48	x	x	x	x			x		x		x
49	x		x	x					x		
50	x		x	x					x		
51	x		x	x					x		x
52	x		x	x					x		x
53	x		x	x					x		x
54	x			x					x		
55	x			x	x			x			
56	x			x	x			x			
57	x			x	x			x			
58	x			x	x			x			
59	x			x	x			x			
60	x			x	x			x			
61	x			x	x			x			
Number active	61	44	26	61	15	15	26	18	15	16	14

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

Control totals were created for each replicate and for each subgroup. Variation was controlled for 10 subgroups, which identified the six SAAL states and the remainder in each census region. The replicated control totals were created by adding error to the current totals for each subgroup. The amount of error depended on the subgroup. The variation comes in the form of the GVF model and uses parameters derived from information in the technical report for the 2003 CPS March Supplement. Table 12-13 provides the GVF parameters for the GVF formula for the standard error of an estimate X related to the NAAL control totals: $SE = \sqrt{AX^2 + BX}$, where B=1,586.

Table 12-13. GVF parameter, by subgroup: 2003

Subgroup	GVF intercept parameter estimate (A)
Rest of Northeast	–0.000008
Massachusetts	–0.000248
New York	–0.000083
Rest of Midwest	–0.000008
Missouri	–0.000284
Rest of South	–0.000008
Kentucky	–0.000395
Maryland	–0.000295
Oklahoma	–0.000464
West	–0.000008

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

The new control totals were created for each replicate *r* as follows:

$$X(r) = X + \frac{z * SE}{\sqrt{60}},$$

where z = a random draw from a standard normal distribution. The error term was divided by the square root of 60 because the stratified jackknife (JK2) was used.

12.1.7.4 Jackknife and Taylor Series Variance Estimation

After the replicate weights had been constructed, the estimate of variance could easily be computed for any statistic. The statistic must be computed 62 times, once using the full-sample weight and an additional 61 times using each of the 61 replicate weights. The variance estimate is the sum of the 61 squared differences between the estimate derived using the full-sample weight and the estimate

derived using each of the 61 replicate weights. That is, the estimate of the variance of a statistic Y is as follows:

$$v(Y) = \sum_{r=1}^{61} (Y_r - Y)^2,$$

where

Y_r = the weighted estimate obtained using the rth replicate weight and

Y = the weighted estimate obtained using the full-sample weight.

The data user can use the variance stratum and variance unit that were developed for the stratified jackknife replicates to compute Taylor series estimates of variance. Taylor series expansion linearizes the estimator and then uses variance estimation methods to estimate the variance of the linearized estimate. The advantage of the linearization method is that it is applicable to general sampling design. However, it requires the derivation of a separate standard error for each nonlinear statistic. In contrast, the jackknife estimator employs a single standard error formula for all statistics.

12.1.7.5 Evaluation of Variance Estimates

Table 12-14 provides standard errors for selected background questionnaire items. The table also shows that the process of replicating control totals, to incorporate the variance associated with CPS estimates discussed in section 12.1.7.3.2, had very little effect on variances computed for the background questionnaire items at the national level. Standard errors were checked at the state level as well, and very little change to the standard errors was observed, as was expected and desired.

Table 12-14. Standard errors for selected background questionnaire items, by background questionnaire item: 2003

Background questionnaire item	Sample size	Estimated percentage	Before replicating control totals		After replicating control totals (final)	
			Standard error	Design effect	Standard error	Design effect
BQ1440: How many hours do you usually watch television, videotapes, or DVDs each day?						
None	378	2.0	0.17	2.53	0.17	2.53
1 hour or less	3,642	22.0	0.48	2.45	0.48	2.45
2 hours	4,970	28.0	0.52	2.48	0.52	2.48
3 hours	3,591	19.6	0.47	2.58	0.47	2.58
4 hours	2,311	12.2	0.32	1.69	0.32	1.70
5 hours	1,146	5.8	0.25	2.09	0.25	2.09
6 or more hours	2,133	10.3	0.43	3.69	0.43	3.69
BQ1620: How often do you read books in English?						
Every day	5,677	31.3	0.47	1.84	0.47	1.84
A few times a week	3,591	20.0	0.44	2.17	0.44	2.17
Once a week	1,868	10.4	0.30	1.79	0.30	1.79
Less than once a week	4,338	25.1	0.49	2.28	0.48	2.27
Never	2,699	13.2	0.36	2.11	0.37	2.11
BQ1865: How much do you think your math skills limit your job opportunities?						
A lot	1,715	9.0	0.39	3.02	0.39	3.02
Some	1,789	10.5	0.44	3.26	0.44	3.27
A little	2,098	12.0	0.44	2.94	0.44	2.95
Not at all	10,572	68.6	0.69	3.61	0.70	3.63
BQ2000: There are 25 or more books in your home right now.						
True	15,195	88.1	0.36	2.28	0.36	2.28
False	2,962	11.9	0.36	2.28	0.36	2.28

See note at end of table.

Table 12-14. Standard errors for selected background questionnaire items, by background questionnaire item: 2003—Continued

Background questionnaire item	Sample size	Estimated percentage	Before replicating control totals		After replicating control totals (final)	
			Standard error	Design effect	Standard error	Design effect
BQ2040: Including yourself, how many people in your house are employed or work for pay or wages?						
None	3,675	16.9	0.40	2.08	0.40	2.08
One	7,099	33.6	0.75	4.63	0.75	4.62
Two	5,372	35.2	0.57	2.63	0.57	2.63
Three or more	1,997	14.3	0.52	3.95	0.52	3.94
BQ2255: In general, how would you rate your overall health?						
Excellent	4,392	25.8	0.56	2.93	0.56	2.93
Very good	6,332	35.7	0.51	2.09	0.51	2.08
Good	4,465	24.5	0.50	2.50	0.50	2.50
Fair	2,172	10.5	0.34	2.24	0.34	2.25
Poor	800	3.5	0.17	1.48	0.17	1.48
BQ2270: Have you ever been diagnosed or identified as having a learning disability?						
Yes	1,006	5.7	0.28	2.58	0.28	2.58
No	17,144	94.3	0.28	2.58	0.28	2.58

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

12.2 WEIGHTING THE PRISON STUDY SAMPLE

The prison study weighting process consisted of four main steps. First, prison base weights were constructed using the prison selection probability (section 12.2.1). Then, a nonresponse adjustment was made to prison base weights to account for nonparticipating prisons (section 12.2.2). Next, inmate base weights were formed using the prison nonresponse-adjusted weight and the within-prison sampling rate (section 12.2.3). Finally, the inmate base weights were raked to control totals to account for inmate nonresponse and noncoverage (section 12.2.4). Section 12.2.5 provides the distribution of the final weights.

Estimates of variance can be made using replication methods (such as the stratified jackknife procedure) or Taylor series linearization. Both methods can take into account the complex sample design of the prison study. To facilitate variance estimation of the prison study outcome measures, stratified jackknife replicate weights were created. The formation of the replicates is described in section 12.2.6, and the resulting variance estimates are evaluated for some background questionnaire items in section 12.2.7. For further information on jackknife and Taylor series variance estimation, refer to section 12.1.7.4.

12.2.1 Computing Prison Base Weights

The prison base weights were computed as the inverse of the prison probability of selection:

$$W_i = \frac{1}{P_i},$$

where

W_i = the base weight for the ith prison and

P_i = the probability of selection for the ith prison.

The distribution of the full-sample base weights for eligible prisons is shown in table 12-15. Because prisons were selected with probability proportional to size, the weights vary with the size of the prison. Table 12-15 also compares the weighted number of inmates with the count of inmates for all prisons on the frame. The weighted number of inmates was calculated using the prison base weight and

the count of inmates in each eligible prison, where the inmate count was updated on the basis of information obtained from the prison contacts immediately prior to data collection.

Table 12-15. Distribution of Prison Study prison base weights: 2003

Prison base weights					Weighted number of inmates	Inmate count from frame
N	Mean	Min	Max	Coefficient of variation		
110	12.75	1.71	84.49	107.95	1,355,833	1,348,458

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

12.2.2 Prison Nonresponse Adjustment

Three prisons did not participate in the prison study. To adjust for the nonparticipating prisons, prison-level variables that are known for both participants and nonparticipants from the frame were used to form nonresponse adjustment cells, and an adjustment factor was applied to participating prisons within each cell. Because all three nonparticipating prisons were state prisons in the Midwest with male inmates only, they were assigned to a cell with eligible prisons of the same region, type of prison, and gender composition. The remaining eligible prisons were assigned to a second nonresponse adjustment cell.

The nonresponse adjustment factor in each cell, F_{α}^{nr}, was computed as the sum of the weighted inmate population for eligible prisons divided by the sum of the weighted inmate population for participating prisons:

$$F_{\alpha}^{nr} = \frac{\sum_{i \in SE(\alpha)} W_i^{base} N_i}{\sum_{i \in SP(\alpha)} W_i^{base} N_i},$$

where

Ni = *the inmate population count of the ith prison;*

$SE(\alpha)$ = *the set of eligible sampled prisons in nonresponse adjustment class* α*; and*

$SP(\alpha)$ = *the set of participating prisons in nonresponse adjustment class* α.

The nonresponse adjustment cells and factors are summarized in table 12-16. The second cell contained no nonparticipating prisons, so the adjustment factor in the cell is equal to 1.000.

Table 12-16. Prison Study prison nonresponse adjustment factors, by adjustment class: 2003

Nonresponse adjustment cell	Eligible prisons		Responding prisons		Nonresponse adjustment factor
	Sample size	Weighted number of inmates	Sample size	Weighted number of inmates	
1: State prisons in the Midwest with male inmates only	21	265,892	18	232,238	1.14
2: Other	89	1,089,941	89	1,089,941	1.00

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

12.2.3 Computing Inmate Base Weights

The full-sample inmate base weight for inmate j of prison i was computed as the product of the prison nonresponse-adjusted base weight and the reciprocal of the inmate sampling rate, as given by

$$W_{ij}^{base} = W_i^{base} F_\alpha^{nr} S_i \frac{1}{P_{ij}},$$

where

S_i = the adjustment factor for the subsampling of units within the ith prison[6] and

P_{ij} = the initial probability of selection for the jth inmate in the ith prison.

The distribution of the inmate full-sample base weights is shown in table 12-17. The variation in the weights is due to the constraint of sampling 9 to 16 inmates per prison and to differences between

[6] One prison was found to have four separate units: one minimum security unit and three reception centers. Because of the difficulty of conducting interviews in all three reception centers, one reception center was sampled from the three with probability proportional to size. The base weights of inmates in the sampled unit were adjusted by the inverse of the selection probability of the unit, S_i. The factor S_i was set to 1 for all other prisons. Inmates were also sampled at a higher rate within the reception center to maintain the same overall selection probability.

inmate counts provided at the time of negotiations with prisons and those determined through the within-prison sampling procedure conducted during data collection.

Table 12-17. Distribution of inmate base weights: 2003

	Inmate base weights				
N	Sum	Mean	Min	Max	Coefficient of variation
1,298	1,358,771	1,047	159	1,423	11

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

12.2.4 Accounting for Nonresponse and Noncoverage

The inmate base weights were raked to the Bureau of Justice Statistics control totals to bring estimates for selected variables to known totals and reduce bias owing to inmate nonresponse and noncoverage. Refer to section 12.1.4.2 for a description of the raking algorithm.

Sampled inmates who completed the background questionnaire were included in the raking process. Literacy-related nonparticipants (those with a language problem or mental disability) were also included because the reasons for nonparticipation are highly related to literacy results. Because raking variables must be nonmissing, the background questionnaire variables of country of birth, educational attainment, and marital status were imputed for inmates who did not complete the background questionnaire for literacy-related reasons. Because there were only 10 cases with missing data, the imputation was done by forming cells on the basis of assessment status code (language problem or mental disability) and then taking the mode of the raking variable in the cell.

Weights were raked to control totals for the following dimensions: region/type of prison, prison security level, inmate gender, race/ethnicity, age category, educational attainment, country of birth, and marital status. The variables were selected in the same manner as discussed in the household weighting process in section 12.1.4.1. The raking factors for each raking dimension category are shown in table 12-18. Raking factors ranged from 0.48 to 1.93. Domains that have a relatively large range of factors will have more variation added to the weights. A maximum of 10 iterations was allowed for the raking process.

After raking, the 3× median rule,[7] discussed in section 12.1.3.3, was used within each region to detect extreme weights. No trimming of the weights was needed.

Table 12-18. Raking factors by raking dimension for the Prison Study, by domain: 2003

Domain	Sample		Control total	Raking factor		
	Size	Estimate		Mean	Minimum	Maximum
Overall	1,173	1,231,421	1,380,776	1.13	0.48	1.93
Region/prison type						
Northeast	137	139,493	174,204	1.25	0.73	1.77
Midwest	209	247,476	242,955	0.98	0.48	1.48
South	485	491,669	530,452	1.08	0.50	1.53
West	206	208,262	273,890	1.32	0.66	1.93
Federal	136	144,521	159,275	1.11	0.57	1.88
Prison security level						
Supermax, max/close/high	375	392,821	466,991	1.19	0.65	1.92
Medium	539	559,789	666,387	1.19	0.69	1.93
Min/low, admin, other	259	278,811	247,398	0.89	0.48	1.38
Gender						
Male	1,097	1,152,051	1,292,354	1.13	0.48	1.93
Female	76	79,370	88,422	1.11	0.61	1.77
Race/ethnicity						
Hispanic	229	237,362	251,137	1.06	0.57	1.66
Non-Hispanic Black only	493	513,458	628,204	1.23	0.55	1.93
Other	451	480,601	501,435	1.05	0.48	1.77
Age						
16–29	391	411,132	513,206	1.26	0.71	1.93
30–49	666	698,107	766,270	1.10	0.61	1.74
50+	116	122,182	101,300	0.83	0.48	1.23
Education						
Less than high school	470	488,881	526,984	1.08	0.48	1.81
High school or higher	703	742,540	853,792	1.16	0.54	1.93
Country of birth						
U.S.	1,051	1,104,383	1,236,811	1.12	0.48	1.93
Other	122	127,038	143,965	1.14	0.74	1.66
Marital status						
Never married	631	661,732	763,735	1.16	0.55	1.92
Other	542	569,690	617,041	1.09	0.48	1.93

NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

[7] The cutoff value of three times the median weight for each cell was used as a guideline to limit the bias introduced by trimming.

12.2.5 Final Inmate Weights

The final inmate weights for inmate j of prison i were computed as the product of the inmate base weight and the raking factor:

$$W_{ij}^{final} = W_{ij}^{base} F_{\beta},$$

where

F_{β} = the raking adjustment factor for raking cell β.

The distribution of the final inmate weights is shown in table 12-19. Note that the raking process increased the coefficient of variation (CV) of the weights from 11.47 percent to 22.63 percent. The overall effect, however, was a reduction in sampling variance. Refer to section 12.2.7 for the evaluation of variance estimates.

12.2.6 Replicate Weights for Variance Estimation for the Prison Study

Because of the clustering of inmates within prisons, simple random sample variance formulas would underestimate sampling variability. Therefore, replicates were formed to facilitate variance estimation. The variance estimation was carried out in three steps: (1) the formation of replicates, (2) the computation of replicate weights, and (3) the estimation of variances and design effects for some survey variables.

To create the replicates, the 110 eligible prisons were sorted in their order of selection. Prisons were paired consecutively and assigned to 55 variance strata. Within each variance stratum, one prison was randomly assigned to variance unit 1 and the other to variance unit 2. Replicates were then formed using the stratified jackknife approach. The rth replicate base weight for the prison associated with variance unit k of variance stratum h was calculated as

$$W_{hk}^{base}(r) = \begin{cases} 0 & \text{if } h = r \text{ and } k = 1 \\ 2W_{hk}^{base} & \text{if } h = r \text{ and } k = 2, \text{ and} \\ W_{hk}^{base} & \text{if } h \neq r, \end{cases}$$

where

r = *1, 2, ..., 55 and*

W_{hk}^{base} = *the full-sample prison base weight for the prison in unit k of variance stratum h.*

Table 12-19. Distribution of Prison Study final inmate weights, by raking dimension: 2003

Domain	Inmate final weights					
	N	Sum	Mean	Min	Max	CV[1]
Overall	1,173	1,380,776	1,177.13	127.11	2,103.34	22.63
Region/prison type						
Northeast	137	174,204	1,271.56	740.22	1,797.64	18.70
Midwest	209	242,955	1,162.46	127.11	2,103.34	28.77
South	485	530,452	1,093.72	505.18	1,583.99	17.76
West	206	273,890	1,329.56	672.21	1,962.41	19.83
Federal	136	159,275	1,171.14	656.27	1,914.11	23.56
Prison security level						
Supermax, max/close/high	375	466,990	1,245.31	663.13	2,103.34	20.22
Medium	539	666,387	1,236.34	127.11	1,962.41	20.35
Min/low, admin, other	259	247,399	955.21	505.18	1,732.04	19.59
Gender						
Male	1,097	1,292,354	1,178.08	127.11	2,103.34	22.83
Female	76	88,422	1,163.45	616.97	1,684.70	19.61
Race/ethnicity						
Hispanic	229	251,137	1,096.67	127.11	1,592.49	19.62
Non-Hispanic Black only	493	628,204	1,274.25	186.37	2,103.34	21.93
Other	451	501,435	1,111.83	151.70	1,695.07	21.69
Age						
16–29	391	513,206	1,312.55	186.37	2,103.34	19.48
30–49	666	766,270	1,150.56	127.11	1,798.30	20.26
50+	116	101,300	873.28	505.18	1,246.99	19.17
Education						
Less than high school	470	526,984	1,121.24	127.11	1,845.37	21.99
High school or higher	703	853,792	1,214.50	151.70	2,103.34	22.46
Country of birth						
U.S.	1,051	1,236,811	1,176.79	127.11	2,103.34	22.93
Other	122	143,965	1,180.04	715.96	1,732.04	19.98
Marital status						
Never married	631	763,735	1,210.36	151.70	1,957.92	20.85
Other	542	617,041	1,138.45	127.11	2,103.34	24.35

[1] Coefficient of variation.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

For each subsequent stage of weighting, adjustments made to the full-sample weights were also made to the replicate weights. As a result of these adjustments, the effect of the weighting procedures was properly reflected in variance estimates. When the stratified jackknife estimation is applied, an estimate of the survey variable is produced using the full-sample weight and is computed 55 additional times using each of the replicate weights.

12.2.7 Evaluation of Variance Estimates

Table 12-20 provides standard errors and design effects for selected variables. Calculations were performed in WesVar (Westat 2002) using stratified jackknife variance estimation. Estimates are for selected variables from the prison study background questionnaire as well as variables used in raking. The table compares design effects using the inmate base weights with those for the final raked weights. The maximum base weight design effect is 2.04. Raking should reduce the variance of variables correlated with the raking variables. After raking, the final design effects of the weights for the raking variables are equal to 0. The maximum design effect of the remaining selected background questionnaire variables is 1.86. In general, due to clustering of inmates within prisons, the design effects are expected to be greater than 1.0. However, due to sampling error associated with variance estimates, some design effects are less than 1.0. For the majority of the selected background questionnaire items, raking resulted in a decrease in variance.

The 55 replicates for the prison study provide sufficient degrees of freedom for stable variance estimates. Table 12-21 shows the active replicates overall and by region and race/ethnicity subdomains.

Table 12-20. Design effects for selected Prison Study background questionnaire and raking variables, by item: 2003

Background questionnaire item	Inmate base weights				Final weights			
	Sample size	Estimated percentage	Standard error	Design effect	Sample size	Estimated percentage	Standard error	Design effect
BQ1495: Before your most recent admission to prison, did you ever serve time in prison, jail, or some other correctional facility as a juvenile or adult?								
Yes	845	73.3	1.46	1.26	845	73.4	1.42	1.19
No	313	26.7	1.46	1.26	313	26.6	1.42	1.19
BQ1635: How much do you think your math skills limit your job opportunities?								
A lot	187	16.4	1.10	1.01	187	16.2	0.98	0.81
Some	213	18.7	1.17	1.03	213	18.7	1.12	0.94
A little	238	20.8	1.31	1.18	238	21.1	1.31	1.17
Not at all	501	44.1	1.65	1.26	501	44.1	1.53	1.08
BQ1680: How many hours do you usually watch television, videotapes, or DVDs each day?								
None	127	10.7	1.02	1.27	127	10.8	1.04	1.30
1 hour or less	271	23.1	1.76	2.02	271	22.7	1.67	1.84
2 hours	244	21.0	1.22	1.05	244	20.6	1.22	1.05
3 hours	176	15.3	1.06	1.00	176	15.7	1.13	1.12
4 hours	127	11.2	0.91	0.98	127	11.1	0.87	0.89
5 hours	75	6.4	0.76	1.10	75	6.6	0.81	1.23
6 or more hours	140	12.3	1.25	1.68	140	12.5	1.28	1.75
BQ1755: How often do you read books in English?								
Every day	583	50.0	1.63	1.24	583	50.2	1.56	1.12
A few times a week	248	21.5	1.38	1.31	248	21.8	1.32	1.19
Once a week	94	8.1	0.86	1.14	94	8.3	0.90	1.24
Less than once a week	135	11.7	0.95	1.01	135	11.5	0.96	1.05
Never	100	8.7	1.11	1.80	100	8.2	0.95	1.39
BQ1785: Have you ever used a computer?								
Yes	292	24.9	1.65	1.69	292	25.0	1.73	1.86
No	869	75.1	1.65	1.69	869	75.0	1.73	1.86

See notes at end of table.

Table 12-20. Design effects for selected Prison Study background questionnaire and raking variables, by item: 2003—Continued

Background questionnaire item	Inmate base weights				Final weights			
	Sample size	Estimated percentage	Standard error	Design effect	Sample size	Estimated percentage	Standard error	Design effect
BQ1945: In general, how would you rate your overall health?								
Excellent	319	27.4	1.62	1.52	319	28.0	1.66	1.59
Very good	399	34.5	1.84	1.74	399	35.0	1.79	1.63
Good	259	22.3	1.24	1.03	259	21.7	1.18	0.96
Fair	132	11.4	0.90	0.93	132	11.1	0.88	0.92
Poor	51	4.5	0.54	0.78	51	4.2	0.51	0.76
BQ1960: Have you ever been diagnosed or identified as having a learning disability?								
Yes	193	16.8	1.17	1.13	193	16.6	1.05	0.91
No	964	83.2	1.17	1.13	964	83.4	1.05	0.91
AGECAT_R: Inmate age category								
16–29	427	33.0	1.87	2.04	391	37.2	0.00	0.00
30–49	736	56.9	1.77	1.65	666	55.5	0.00	0.00
50+	130	10.1	0.97	1.34	116	7.3	0.00	0.00
BQRACETH_RAKE: Inmate race/ethnicity								
Hispanic	245	18.8	1.44	1.75	229	18.2	0.00	0.00
Non-Hispanic Black only	539	41.3	1.67	1.49	493	45.5	0.00	0.00
Other	508	39.9	1.86	1.87	451	36.3	0.00	0.00
EDUCCAT_R: Inmate highest level of education								
Less than high school	458	39.1	1.64	1.31	470	38.2	0.00	0.00
High school or higher	703	60.9	1.64	1.31	703	61.8	0.00	0.00

NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Table 12-21. Prison Study active replicates, by selected subdomains: 2003

Replicate	Total	Region				Race/ethnicity		
		Northeast	Midwest	South	West	Hispanic	Non-Hispanic Black only	Other
1	x	x				x	x	x
2	x	x				x	x	x
3	x	x				x	x	x
4	x	x				x	x	x
5	x	x				x	x	x
6	x	x				x	x	x
7	x	x				x	x	x
8	x	x				x	x	x
9	x		x				x	x
10	x		x			x	x	x
11	x		x				x	x
12	x		x			x	x	x
13	x		x				x	x
14	x		x			x	x	x
15	x		x			x	x	x
16	x		x			x	x	x
17	x		x			x	x	x
18	x		x			x	x	x
19	x		x			x	x	x
20	x			x		x	x	x
21	x			x		x	x	x
22	x			x		x	x	x
23	x			x		x	x	x
24	x			x		x	x	x
25	x			x		x	x	x
26	x			x		x	x	x
27	x			x		x	x	x
28	x			x		x	x	x
29	x			x			x	x
30	x			x		x	x	x
31	x			x		x	x	x
32	x			x		x	x	x
33	x			x		x	x	x
34	x			x		x	x	x

See notes at end of table.

Table 12-21. Prison Study active replicates, by select subdomains: 2003—Continued

Replicate	Total	Region				Race/ethnicity		
		Northeast	Midwest	South	West	Hispanic	Non-Hispanic Black only	Other
35	x			x		x	x	x
36	x			x		x	x	x
37	x			x		x	x	x
38	x			x			x	x
39	x			x		x	x	x
40	x			x			x	x
41	x			x		x	x	x
42	x			x		x	x	x
43	x			x		x	x	x
44	x				x	x	x	x
45	x				x	x	x	x
46	x				x	x	x	x
47	x				x	x	x	x
48	x				x	x	x	x
49	x				x	x	x	x
50	x				x	x	x	x
51	x				x	x	x	x
52	x				x	x	x	x
53	x				x	x	x	x
54	x				x	x	x	x
55	x				x	x	x	x
No. active	55	8	11	24	12	49	55	55

NOTE: All adults of Hispanic origin are classified as Hispanic regardless of race. Those classified as Black are non-Hispanic Black only. Those classified as other include non-Hispanics of all other races including multiracial. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

This page is intentionally left blank.

CHAPTER 13

SCORING

Justin Baer, American Institutes for Research, Jared Bernstein, Ordinate Corporation,
and Michelle Amsbary, Westat

This chapter describes the procedures followed for scoring the three main components of the 2003 assessment: the cognitive items, the Fluency Addition to NAAL (FAN), and the Adult Literacy Supplemental Assessment (ALSA). For the cognitive items and the ALSA, the scoring procedures used were similar to the procedures implemented for scoring the NAAL field test (chapter 4). Scoring the FAN was more complex because the scores were generated by an automatic speech recognition (ASR) system. To ensure the validity of the FAN data, a sample of tasks scored by the ASR were compared to a sample of tasks scored by human scorers.

13.1 COGNITIVE ITEMS

13.1.1 Refinement of Training Materials

The scorer training materials for the 2003 main study assessment were largely adapted from those used for the NAAL field test (see chapter 4). Prior to the scoring of the NAAL field test, AIR staff compiled sample responses to each of the cognitive items in the field-test assessment booklets. In developing the training materials for the 2003 assessment, AIR staff began by reviewing the inter-rater reliability statistics for the field-test items selected for the main study.[1] Sample responses to items with high inter-rater reliability during the field test were reused as scorer training materials. For items with low inter-rater reliability and those that were challenging to score, AIR conducted additional range finding to locate both straightforward and ambiguous responses. During range finding, AIR staff reviewed a sample of booklets that had been returned from the field to the data collection facility, searching for responses that would be valuable for training purposes.

Sample responses to the targeted items were then photocopied and compiled with the existing sample responses to items from the field test with high inter-rater reliability. The complete collection of sample papers consisted of a mixture of responses that closely matched the scoring rubrics and more ambiguous responses to items. Because the sample responses were to be used as scorer training papers, the number of example papers selected also varied across the items on the basis of their scoring difficulty.

[1] See chapter 4 (section 4.2.2.2) for a discussion of procedures for items with low inter-rater reliability. Inter-rater reliability statistics for field test items are also presented in chapter 4.

For example, fewer examples of responses to quantitative questions were photocopied because correct answers were typically numeric and simple to score.

As noted in chapter 2, 6 blocks of items from the 1992 National Adult Literacy Study (NALS) were included in the 2003 NAAL. When items are re-used across assessments, trend scoring is usually conducted to ensure that the common items are scored consistently from one assessment to the other. Unfortunately, with the exception of the scoring rubrics, none of the training material from the 1992 survey was available. Consequently, trend scoring between the 1992 NALS and the 2003 NAAL could not be conducted. To ensure consistency in scoring, AIR staff consulted with a project member from the 1992 survey who reviewed training papers for the 1992 items reused in 2003. Her comments and interpretation of the 1992 scoring rubrics were carefully documented so they could be included in the materials used to the train the scorers.

For the new blocks developed for the 2003 assessment, AIR convened an expert panel to review the scores assigned to the training papers.[2] The panel was directed to closely review the rubrics and training papers for items with low inter-rater reliability in the field test. This meeting also provided a final opportunity to review and modify the scoring rubrics for the new 2003 blocks. On the basis of decisions made by the expert panel, several of the scoring rubrics were revised further. Comments from the panel about how responses to particular assessment items should be scored were also documented and included in the scorer training materials.

13.1.2 Scorer Training

Once the scores assigned to the training papers had been reviewed and agreed on by the members of the expert panel, the papers were compiled in training binders. For each item, one or two training papers were designated anchor papers. Anchor papers were straightforward responses to assessment questions that clearly corresponded to the rubrics. The remaining training papers were a mix of straightforward and more challenging responses designed to expose scorers to the range of responses they might encounter and to make certain that they demonstrated sufficient aptitude to score the assessment.

Scoring was conducted at the scoring contractor's facility in Tucson, Arizona. A total of 142 scorers were hired to score the exercise booklets; all scorers were required to hold a bachelor's degree from a college or university. Each scorer was assigned to a table that was responsible for scoring a

[2] The panelists had expertise in reading instruction and assessment as well as the alignment of curriculum, instruction, and assessment.

specific assessment block, for a total of 13 tables with 9 to 13 scorers per table. A scoring supervisor with previous scoring experience was also assigned to each table to answer questions and monitor the scoring. The 13 scoring supervisors all had previous experience working at the scoring center and had demonstrated proficiency scoring assessments. AIR staff trained the scoring supervisors, who in turn trained the scorers.

To guide the scorer training, each item was accompanied by a script written to incorporate a discussion of the items as well as a review of the training papers. The script for each item described the purpose of the question and explained how the item should be scored. For items for which a correct answer could be expressed a variety of ways, the scripts advised the scorers that the rubrics and training papers did not include every possible correct response. For these items, the scorers were instructed to use the rubrics and training papers as a guide for determining whether a response should be scored as correct.

Scorers were given time to read the item rubrics before reviewing and discussing the anchor papers as a group. Following the review, the scorers independently scored the training papers, which were then discussed by all the scorers assigned to a particular block. NAAL staff monitored the training to make certain that the scorers correctly scored the training papers. NAAL staff remained at the scoring center to answer questions during scorer training and through the first two days of scoring and were available by phone for consultation until scoring concluded.

13.1.3 Scoring Procedures and Quality Control

Scoring of the household assessment began in late January 2004 and was completed by mid-February 2004. The assessment booklets were rotated among the tables of scorers, with each table scoring its assigned block and then passing the booklets along to another table for scoring. To ensure consistency in scoring, half the booklets were rescored by two different scorers. In addition, the scoring supervisor for each block "backread" approximately 10 percent of all item responses scored by each rater. During "backreading," the scoring supervisor compared the scores awarded by one rater to the scorers awarded by a second rater to the same booklet. The scoring supervisor then discussed any discrepancies with the individual scorers as well as with the other scorers at the table.

Following the end of scoring for each day, the scoring contractor generated inter-rater reliability statistics for the sample of rescored booklets. Reliability was measured as the frequency of agreements between two scorers for each item scored. NAAL staff reviewed the reliability statistics daily for each assessment item to make certain that the rubrics were applied consistently across scorers. The number of possible score points and final inter-rater reliability for each item are summarized in table 13-1. The

number of score points per item ranged from 2 to 10. Of the 153 items, all but 3 had inter-rater reliability agreement greater than 95%.[3] The average reliability of the three remaining items was 94%. For the entire pool of items in the assessment, the final inter-rater reliability across all items in the household and prison samples, including the core, was 99%. Inter-rater reliability ranged from a low of 92.6% (item N011101) to a high of 100% (items CC001, C030601, and C060101).

In addition to scoring the 13 blocks of assessment items, the scorers also scored the 7 core items originally scored in the field by the interviewers. In the field, performance on the core was used to assign respondents to either the main assessment (NAAL) or the supplementary assessment (ALSA). Although the determination of whether a respondent should be assigned to NAAL or to ALSA had already been made in the field, the NAAL scorers rescored the core items in each assessment booklet. As noted in table 13-1, the inter-rater agreement between the scorers for each of the 7 items was greater than 99%. The core scores from the NAAL scorers, rather than the interviewer core scores, were used when the data were scaled and proficiency scores were generated. Core scores from the NAAL scorers were used because scoring consistency could be assessed through the inter-rater reliability statistics calculated for all items scored by the scoring contractor staff.

Table 13-1. Inter-rater reliability statistics for NAAL household and prison items, by block: 2003

Block	Item	Score points	Reliability
Core	CC001	2	100.0
	CC002	2	99.7
	CC003	2	99.4
	CC004	2	99.5
	CC005	2	99.6
	CC006	2	99.6
	CC007	2	99.6
Block 1	N010101	2	98.6
	N010201	2	98.6
	N010301	2	96.9
	N010401	2	99.6
	N010501	3	98.1
	N010601	2	97.1
	N010701	3	97.1
	N010801	2	98.1
	N010901	2	97.0
	N011001	4	98.0
	N011101	10	92.6
Block 2	C020101	2	98.4
	C020201	2	98.3
	C020301	5	99.0

See notes at end of table.

[3] This count is based on rounding the inter-rater reliability statistics to the nearest integer.

Table 13-1. Inter-rater reliability statistics for NAAL household and prison items, by block: 2003—Continued

Block	Item	Score points	Reliability
	C020401	2	98.1
	C020501	2	98.0
	C020601	3	99.0
	C020701	6	98.3
	C020801	8	98.2
	C020901	3	97.3
	C021001	2	99.0
	C021101	2	98.6
Block 3	C030101	3	97.2
	C030201	2	99.9
	C030301	3	97.3
	C030401	2	99.9
	C030501	3	98.9
	C030601	2	100.0
	C030701	3	98.1
	C030702	3	97.3
	C030703	3	99.6
	C030704	3	99.1
	C030705	3	98.9
	C030706	4	96.0
	C030707	3	99.6
	C030708	3	99.4
	C030709	3	98.7
Block 4	C040101	2	99.8
	C040201	2	99.9
	C040301	2	97.2
	C040401	4	98.4
	C040501	4	94.4
	C040502	2	99.2
	C040503	3	98.3
	C040504	2	98.9
	C040601	2	98.8
	C040701	3	96.0
	C040801	6	98.0
Block 5	C050101	2	99.7
	C050201	3	99.0
	C050301	3	99.3
	C050401	3	98.4
	C050501	2	99.4
	C050601	4	99.0

See notes at end of table.

Table 13-1. Inter-rater reliability statistics for NAAL household and prison items, by block: 2003—Continued

Block	Item	Score points	Reliability
	C050701	3	98.9
	C050801	3	99.2
	C050901	2	94.6
	C051001	2	94.7
	C051101	2	96.6
Block 6	C060101	2	100.0
	C060201	2	99.3
	C060301	4	99.4
	C060401	3	99.3
	C060501	2	99.3
	C060601	2	99.5
	C060701	6	99.4
	C060801	4	99.6
	C060901	2	97.2
	C061001	3	98.9
	C061101	4	97.3
Block 7	C070101	2	99.7
	C070201	2	99.2
	C070301	4	99.4
	C070401	2	99.1
	C070501	2	99.8
	C070601	4	99.6
	C070701	3	98.9
	C070801	6	99.5
	C070901	2	98.9
	C071001	2	99.8
	C071101	3	97.4
Block 8	C080101	3	98.9
	C080201	2	99.0
	C080301	2	99.2
	C080401	2	99.1
	C080501	2	99.8
	C080502	3	99.4
	C080503	3	99.4
	C080504	2	99.8
	C080601	2	99.6
	C080701	4	98.2
	C080801	2	99.6

See notes at end of table.

Table 13-1. Inter-rater reliability statistics for NAAL household and prison items, by block: 2003—Continued

Block	Item	Score points	Reliability
Block 9	N090101	3	98.2
	N090201	3	97.0
	N090301	2	97.7
	N090401	2	98.0
	N090501	2	97.5
	N090601	2	98.3
	N090701	2	99.5
	N090801	2	97.7
	N090901	3	95.7
	N091001	5	98.6
Block 10	N100101	2	99.9
	N100201	2	99.9
	N100301	2	99.9
	N100401	3	97.1
	N100501	2	99.6
	N100601	3	99.6
	N100701	3	97.5
	N100801	2	99.8
	N100901	2	99.8
	N101001	2	99.8
Block 11	N110101	2	98.9
	N110201	2	98.7
	N110301	4	97.6
	N110302	5	99.2
	N110303	6	98.7
	N110401	3	98.1
	N110501	2	98.8
	N110601	3	97.4
	N110701	2	99.5
	N110801	3	99.2
	N110901	3	96.2
Block 12	N120101	2	98.3
	N120201	2	98.8
	N120301	2	97.8
	N120401	3	93.5
	N120501	2	97.9
	N120601	3	99.6
	N120701	4	99.0
	N120801	4	98.4

See notes at end of table.

Table 13-1. Inter-rater reliability statistics for NAAL household and prison items, by block: 2003—Continued

Block	Item	Score points	Reliability
	N120901	2	99.7
	N121001	3	97.9
	N121101	4	98.5
Block 13	N130101	5	98.0
	N130102	4	99.6
	N130103	5	99.8
	N130104	5	99.6
	N130201	2	99.6
	N130301	4	97.2
	N130401	2	99.0
	N130501	6	99.5
	N130601	4	99.5
	N130701	6	95.7
	N130801	2	99.8
	CN130901	6	99.4
Mean	†	3	98.6

†Not applicable.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

13.1.4 Scoring the Prison Sample

Although the scoring of the household assessment was completed by February 2004, data collection for the prison sample continued through July 2004. Because the household and prison assessments could not be scored concurrently, special provisions were implemented to make certain that the prison sample was scored consistently with the household sample. High scoring reliability between the two samples was especially important because the household and prison assessments were to be combined into a single reporting sample.

To maintain consistency in scoring, the same scorers recruited for the household sample were hired to score the prison sample and were assigned to the same block they had scored previously. Prior to the scoring of the prison booklets, the scoring supervisors for each block spent several hours reviewing the training materials with their table. Additionally, a sample of 882 household booklets scored earlier in the year was drawn for rescoring. The 882 household booklets were divided equally among the 26 combinations of assessment items, for a total of approximately 34 booklets for each combination of items. The booklets selected were stratified on the basis of their total booklet score, ensuring that the rescoring sample comprised assessments from respondents with varying levels of literacy.

The booklets were distributed to the scoring tables where each table scored its assigned block. Once each booklet had been scored, the scores assigned during the rescoring were compared with the original scores assigned in January and February 2004. Consistency in scoring was evaluated by examining the frequency with which the second scores agreed with the first scores, similar to an inter-rater reliability statistic. NAAL staff reviewed the statistics for each item and discussed items with low agreement (less than 95%) with the scorers. Once the review of training materials was completed, scoring began for the prison sample. Similar to the household sample, half of all assessment booklets were rescored to monitor inter-rater scorer agreement. With the exception of booklets from California prisons (discussed below), all prison assessments were scored by August 2004.

Assessments from California prisons could not be scored concurrently with those from other prisons because data collection in California was extended into the summer of 2004. Because all other assessments for the prison sample had already been scored by August 2004, the sample of 116 assessments from California prisons was scored directly by the two AIR staff members who developed the scoring materials and supervised the household and prison scoring.

AIR staff followed procedures for scoring the California assessments that were similar to those implemented to score the other assessments in the prison sample. First, to ensure consistency in scoring, AIR scorers independently rescored a sample of booklets and compared their scores against each other. The California prison assessments were then scored, with half the booklets randomly selected to be rescored to check the inter-rater scoring reliability. On completion of scoring, the California assessments were then combined with the remaining prison sample booklets. Scoring of the California prison assessments was completed in September 2004.

13.2 FLUENCY ADDITION TO NAAL (FAN)

13.2.1 Background

To evaluate the level of oral reading fluency of NAAL participants, the Fluency Addition to NAAL (FAN) was administered. For the FAN, each respondent read aloud from lists and passages of text.[4] The oral reading responses were digitally recorded and subsequently analyzed for measures of accuracy and fluency (accurate reading rate). Each list and each passage was digitized and saved to a separate audio file referred to as a response recording. The response recordings were then sent to a scoring contractor for machine scoring. Due to technical difficulties with the FAN software and

[4] Additional information about the background and design of the FAN is presented in chapter 2.

associated hardware, as well as transmission issues between the data collection and scoring contractors, there were approximately 40 cases for which primary machine scores are not available.

13.2.2 Scoring the FAN Data

To automatically score the FAN responses, each respondent's oral reading of the FAN material was first digitally recorded during the FAN portion of the NAAL assessment. Respondents read into a microphone that was connected to a laptop computer. Special audio recording software from the contractor for the fluency assessment was installed on each laptop and allowed the audio from the microphone to be digitized and recorded. The response recordings were then downloaded from the computer and sent to the contractor for automatic scoring.

The first step in scoring the FAN recordings using an automatic speech recognition (ASR) system was the development of a language model rule set. As part of the language model training, professional transcribers were hired to transcribe about 1,000 responses for each target item in the FAN and also all data used to validate the ASR system. A transcription is a string of words and symbols that represent the recorded response of the respondent. From these transcriptions, words were identified that were not already in the dictionary of the FAN contractor's speech recognition system. Pronunciations were created for missing words and were inserted into the dictionary.

The transcriptions were then divided into two sets: a training set of 4,681 responses and a test set of 2,170 responses. The training set was used to build the language model rule set and the test set was used to test it. The two sets did not intersect. The transcriptions were tagged for part of speech. For example, the word *dog* was tagged as a singular noun and the word *the* was tagged as a determiner. The actual tags used in this process were from the well-established Penn Treebank Tag Set (Marcus et al. 1993) and were rich in grammatical information. The tagging was done such that specific word-level rules, such as *trained* goes to *train* and *visited* goes to *visit*, could be generalized as "Verb-ed goes to Verb."

After the tagging was completed, a preliminary language model was created for each passage. (The language model is specific to each passage or list). This preliminary language model consisted of the answer choice or the text of the passage, for example, the string of words "Curly is my big black dog …" and a small set of rules. These rules were written by linguists and were intended to account for as many reading errors and disfluencies as possible. Then all the transcriptions for the passage were run through this language model. When the transcription found a path through the language model, an output file recorded any rules that fired so that the information could be reviewed by a linguist and modified if

necessary. Specifically, more descriptive rules, such as "Noun goes to Noun-s," were added to account for more reading errors. If a transcription could not be matched to a path through the language model, a new rule was written to account for it. This process was iterated until the number of transcriptions accounted for was maximized.

After the language models were created, the responses were machine scored. The speech recognition engine was used to identify the string of words that best matched the speech in each response recording. An alignment algorithm was then applied to the string of words to determine the number of omissions, substitutions, and insertions. A subset of respondents completed the Adult Literacy Supplemental Assessment (ALSA)[5] instead of the main NAAL study. For ALSA respondents, all the response recordings were transcribed, and the alignment algorithm was applied to the human transcriptions instead of to the output of the speech recognizer. On the basis of a scoring algorithm, the number of reading errors was tallied and weighted to produce the number of words read correctly for each response recording.

To evaluate the validity of the scores generated by the ASR system, a sample of scores from the ASR system was compared to a sample of scores from human scorers. The sample was comprised of recordings from 480 respondents. To ensure the validity of the ASR system across key population groups, the sample was stratified by performance on the NAAL cognitive tasks (as well as respondents who completed the Adult Literacy Supplemental Assessment) and by the following linguistic/ethnic groups: Black adults, Spanish-speaking adults, and Other English-speaking adults. The validity analyses were performed on each of the passage included in the FAN as well as all the three English word lists. The final correlations between the ASR system scoring and the human scoring of the same tasks are presented in table 13-2.

[5]Respondents were administered either the main NAAL survey or the ALSA. The decision was based on the respondent's performance on a set of screening items. The ALSA used concrete stimulus materials and visual input to support the assessment of the least-literate adults.

Table 13-2. Correlations between human ratings and machine scores of number of words read correctly, by Fluency Addition to the NAAL (FAN) task: 2003

Task	Correlation
Passages	
1st Grade 3 passage	.96
2nd Grade 3 passage	.98
3rd Grade 3 passage	.99
1st Grade 8 passage	.98
4th Grade 3 passage	1.00
2nd Grade 8 passage	1.00
3rd Grade 8 passage	.99
4th Grade 8 passage	1.00
Word lists	
1	.98
2	.99
3	.98

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

13.2.3 FAN Data Products

Following the completion of scoring, six primary data products were generated for each response recording:

1. Span summary
   - index of first word attempted (first)
   - index of last word attempted (last)
   - number of words read correctly (nwordcorr)
   - narrow time in centiseconds (narrowt)
2. Articulation rate (pros)
3. Number of short pauses (nsp)
4. Number of long pauses (nlp)
5. Number of words completely deleted (ndel)
6. Number of false starts (fstart)

13.2.3.1 Span Summary

In the source text, each word has an index value that corresponds to its sequential position in the text. The first word is 0, the next word is 1, the next is 2, and so on. The index values of the first word attempted and the last word attempted by the respondent are the first two items in the span summary.

The number of words read correctly is the third item in the span summary. This number is an estimate of the number of words the respondent read correctly in the source text and is optimized to match human ratings of reading accuracy.

The last number in the span summary is narrow time. Narrow time is the time from the onset of the first word spoken in the response recording through the offset of the last word spoken. The value is in centiseconds (e.g., 6000 centiseconds equals one minute).

Note that only the span summary was analyzed as part of the validation of the ASR system. The other measures listed below are for research purposes only.

13.2.3.2 Articulation Rate

The articulation rate, or phonemic rate of speech (pros), is defined as the number of phonemes per second of speech. The articulation rate is computed by counting the number of phonemes in the response and dividing by the total speech duration. Total speech duration is the sum of the elapsed time of the relevant spoken material and does not include inter-word pauses or leading or trailing silence in the response file. For the purposes of computing pros, all filler material (e.g., hesitations, mouth noise) is treated as a part of inter-word pauses and is not included in the total duration of speech.

13.2.3.3 Number of Short Pauses

The number of short pauses (nsp) is the count of pauses with durations that are greater than 200ms but less than 1000ms. The pause duration is a measurement of the time between the speech sounds of two contiguous words. The pauses at sentence boundaries are treated the same as pauses elsewhere in the respondent's speech.

13.2.3.4 Number of Long Pauses

The number of long pauses (nlp) is the count of pauses with durations equal to or greater than 1000ms. The duration is computed from the end of one word to the beginning of the next. As with nsp, pauses at sentence boundaries are treated the same as all other pauses.

13.2.3.5 Number of Deletions

The number of deletions (ndel) is the number of words completely omitted in the oral reading response (between the first word and last word attempted) in addition to the words that were deleted because of a substitution.

The process of aligning the most likely hypothesis of what the respondent said with the source text produces an estimate of the number of deletions, insertions, and substitutions present in the word sequence. From this estimate, the number of deletions is added to the number of substitutions to produce the total ndel value.

13.2.3.6 False Start Count

The false start count (fstart) is the count of how many times a respondent backed up in the text to repeat (or attempt to repeat) previous words. Even if the text was read incorrectly during either the first attempt or the repeat, the event is identified as a false start as long as there is enough correct content to not categorize the event as a substitution or insertion of irrelevant material.

The false start count is extracted from the path through the language models. The language model for each FAN task encompasses information about the errors and disfluencies that a respondent is likely to make when reading a text. These errors and disfluencies include such things as substitutions, insertions, and false starts, and they are referred to as "rule firings." When the FAN contractor's system determines the best path through the language model that matches a respondent's utterance, the system also tracks which rules are fired. The false start count is the number of times the false start rule fired in the language model to traverse the best path for a given response recording.

13.2.3.7 Secondary Machine Scores

Most language models for passages contain many rules. As with the false start count, these rules can provide information about the type of disfluencies and errors that are made as a respondent reads a

passage. The secondary machine scores simply list the number of times each of the most frequently Occurring Rules Was Fired For A Given Response Recording Of A Passage Reading.

13.3 ADULT LITERACY SUPPLEMENTAL ASSESSMENT

13.3.1 Background and Rationale

The Adult Literacy Supplemental Assessment (ALSA) was developed as an alternative assessment to the main exercise used in the National Assessment of Adult Literacy (NAAL) in an effort to gather as much information as possible about adults with limited English literacy skills. (See section 2.6 for a complete discussion of the development and content of ALSA.) On the basis of a respondent's performance on the seven core items in the exercise booklet, an algorithm in the computer-assisted personal interviewing (CAPI) system was implemented to determine whether the respondent should continue with the main exercise or be directed to ALSA. The ALSA tasks allowed adults with limited literacy skills to demonstrate their abilities to understand and use printed materials in ways that the traditional NAAL exercise items do not.

The algorithm used to determine whether respondents should complete the main exercise tasks or ALSA was incorporated into the CAPI interviewer guide. Interviewers were trained to score the core assessment items (as explained in section 13.3.2) immediately after the respondent completed the items. On the basis of the interviewer's evaluation of whether the core item responses were correct, incorrect, or not provided, the CAPI system indicated which assessment to administer. The algorithm took into account the response to the core item and the language in which the core items had been administered (English or Spanish). The algorithm directed the respondent to ALSA under three scenarios:

- Core items CC001 through CC006 were all wrong or were not answered, and the items were administered in English.
- Core item CC007 was wrong or was not answered and was administered in English, and no answer was provided for core items CC003 and CC004.
- The core items were administered in Spanish, fewer than five of core items CC001 through CC006 were correct, and core item CC007 was wrong or not answered.

Under all other circumstances, the interviewer was instructed to continue with the main assessment booklet tasks.

13.3.2 Field Scoring of Core Items to Identify Respondents for the Adult Literacy Supplemental Assessment

As mentioned above, responses to the seven core items were scored in the field by the interviewers. Immediately following the completion of these seven items, the CAPI interviewer guide instructed the interviewer to take the assessment booklet from the respondent. The interviewer guide then led the interviewer through the scoring process for each core item. For each task, the interviewer guide provided the interviewer with the acceptable response(s). After a quick review of the respondent's answer, the interviewer determined whether it was acceptable, was not acceptable, or had been left blank and entered this information into the CAPI system.

On the basis of the interviewer's determination of the core responses, and the language in which the core items were administered, the algorithm was invoked to determine whether the interviewer should administer ALSA or the main exercise.

13.3.2.1 Interviewer Training

Three main household study training sessions were devoted to the scoring of the core items. The scoring determined whether NAAL or ALSA would be administered and was a focus of the assessment training sessions. It was crucial that interviewers apply the scoring guidelines uniformly to ensure that respondents were not routed to the wrong assessment.

Included in the core scoring training session were interactive exercises with examples, a thorough discussion of correct and incorrect answers to each core item, and practice exercises that used actual responses from the 2001 NAAL field-test booklets.

Interviewers were trained to give respondents the benefit of the doubt while still following the scoring guidelines in the interviewer guide. It was understood that the scoring rules in the interviewer guide could not anticipate every possible answer found in the field.

13.3.2.2 Quality Control

As part of the standard NAAL quality control procedures, all seven core items in the completed assessment booklets returned to the home office of the data collection contractor were rescored by trained

home office staff members. The validation results were entered into a specially designed core scoring program. The program compared the interviewer's scoring with that of the home office staff, enabling supervisors to provide interviewers with feedback on their performance.

Early in the data collection period, the core items were rescored for 100 percent of the receipted assessment booklets. When interviewers were determined to be proficient at scoring the core items (based on an 85 percent match between the interviewer and the home office staff), no further core validation was conducted for those interviewers. Home office staff continued to rescore 100 percent of the core items for the other interviewers.

The in-house validation of the core items continued through the middle of January 2004, as the end of data collection approached. As table 13-3 shows, home office staff rescored a total of 13,608 core assessments. Of the rescored core assessments, only 115 discrepancies (0.85 percent) between interviewer and home office scores resulted in misclassification; that is, the rescoring effort assigned the case to a different ALSA/main assessment route than the one determined through the interviewer's scoring. Therefore, although there was significant disagreement between the home office and interviewer scoring (19 percent), the respondent actually completed the incorrect assessment in fewer than 1 percent of these cases.

Core item CC004 produced the largest number of scoring discrepancies between the home office staff and the interviewers – slightly more than 1,000 discrepancies. The item required the respondents to underline a sentence in the 5-paragraph long stimulus material. There were two possible correct sentences, as well as text in both sentences that was optional. This ambiguity likely led to discrepancies in the application of the scoring rubrics.

13.3.3 Recording of Responses

The administration of ALSA required a higher level of interaction between the interviewer and the respondent than did the main assessment booklet. For the main assessment, interviewers were responsible only for guiding the respondents through the items. However, for ALSA, the interviewers read each question to the respondent and classified the response.

Table 13-3. Summary of discrepancies in core item scoring between interviewers and home office staff, by core item: 2004

Core item	Number
Number of core assessments rescored	13,608
Number (%) of discrepancies	2,534 (19%)
Number of discrepancies by item number	
CC001	169
CC002	371
CC003	690
CC004	1,057
CC005	324
CC006	464
CC007	444
Number (%) of discrepancies resulting in misclassification	115 (0.85%)
Rescored as ALSA	49
Rescored as NAAL	66

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

ALSA was designed so that the response categories would be easy to find and the classification rules easy to follow. Response categories for each item were enclosed in a box. Each response box contained a question to help interviewers determine which response classification to select, such as "WHAT DID SP SAY?" The interviewer selected the response category that most accurately reflected the respondent's answer or action in response to each item.

13.3.3.1 Interviewer Training

Interviewers were given extensive training in the administration of the ALSA instrument. A training DVD was developed to ensure that all interviewers received the same standardized training. Training concentrated on how to accurately classify respondent answers, follow skip patterns, use nondirective probing techniques, and gain cooperation, as well as the appropriate use of stimulus materials.

As part of a certification exercise at the end of the DVD, interviewers listened to a respondent provide an answer and then classified it on the questionnaire. This exercise was collected during training, and the results were reviewed with the interviewers.

Because the ALSA interview was administered in a small sample of cases, most interviewers did not administer it on a regular basis. Therefore, the interviewers were required to review the training DVD throughout the data collection period as refresher training.

13.3.3.2 Quality Control

The ALSA booklets were receipted, reviewed, and edited in the home office. Trained staff reviewed notes written in the margins and ensured that every questionnaire item had a valid response and that the skip patterns had been followed correctly by the interviewer.

13.3.3.3 Data Entry of Questionnaires

Completed ALSA booklets were entered by the data collection contractor's data entry staff. Once all responses had been keyed, the codebooks and frequencies generated from the data were reviewed for accuracy and completeness and then reconciled. The final, clean data set was submitted to the American Institutes for Research (AIR) for analysis at the end of the field period.

This page is intentionally left blank.

CHAPTER 14

ITEM ANALYSIS, SCALING, AND ESTIMATES OF SUBPOPULATION PROFICIENCIES

Stéphane Baldi and Justin Baer, American Institutes for Research[1]

The National Assessment of Adult Literacy (NAAL) scales are reported on the same three proficiency scales—prose, document, and quantitative—used for the 1992 National Adult Literacy Survey (NALS). This chapter describes the procedures and models used to conduct item analysis, scale the results, estimate respondents' proficiency, and conduct statistical analyses.

14.1 PROCEDURES AND QUALITY CONTROL

After the assessment booklets were scored by a contractor, the scored data were sent to NAAL staff for item analysis and scaling. To ensure the accuracy of the item analysis and scaling, NAAL staff implemented two key quality control steps. First, as described below, all analyses were conducted by two independent teams of NAAL staff. Second, all scaling activities were performed using two software packages, PARSCALE (Muraki and Bock 2003) and AM (Cohen et al. 2000). PARSCALE is a software package capable of performing item response theory (IRT) scaling and scoring or rating scale data. AM is a statistical software package capable of IRT scaling and analyzing data from complex samples, especially large-scale assessment data such as the NAAL. Using both PARSCALE and AM allowed the analysis staff to check the reliability of the estimated item parameters and to make certain that the estimates were consistent regardless of the software package employed.

14.1.1 Analysis Teams

To provide independent verification of the results, all item analysis and scaling tasks were completed by two independent teams of analysts. Within the teams, analysts were allowed to consult with one another and compare results. For the item analysis, the two teams ran their analyses and then submitted them to a research assistant to compare the results. The research assistant flagged any inconsistencies between the two sets of analyses, which were then resolved in joint team meetings with senior project staff. The final set of item analysis statistics was verified by both analysis teams.

[1] Eugene Johnson contributed material to an early draft of this chapter.

For item analysis and scaling, one team scaled the assessment items using PARSCALE, while the other scaled the items using AM. The analysis teams and senior project staff jointly reviewed the item statistics and item parameters generated by PARSCALE and AM to ensure consistency. In the end, the item parameters generated by AM were used instead of those generated by PARSCALE, because of AM's greater precision in estimating standard errors.

14.2 ITEM ANALYSIS

NAAL staff calculated p-values and adjusted biserial/polyserial correlations between an item and the total booklet score in which the item appeared (with the item excluded from the score). P-values of the NAAL items can be found in appendix E. Following the quality control steps, all item analyses were conducted by two independent research teams.

On the basis of the examination of the item analysis statistics, one item, C060401, was dropped prior to scaling. Because of an error in the reproduction of the stimulus material accompanying the item, the item became much more difficult to interpret. A substantial number of respondents skipped the item (11%), far more than the number of respondents who omitted the preceding and succeeding items (5% and 4%, respectively). Further, the difficulty of the item increased from the field test, where the stimulus material was properly formatted. The analysis staff concluded that the formatting problem with the stimulus adversely affected respondents' ability to answer the question and decided to drop it from scaling. The item analysis revealed no problems with the remaining 152 items, which were retained for scaling.

14.2.1 Partial Credit Items

Although partial credit points were collected for some items in 1992, partial credit was not awarded when the items were scaled. Prior to scaling the 2003 survey data, NAAL staff reviewed the common 1992 and 2003 items to determine whether partial credit could be awarded to the items for which partial credit points were collected. In reviewing the items, analysts followed the same rules used to assign partial credit for the 2003 items:

- A score point assigned partial credit must substantively make sense as partial credit. In other words, to receive partial credit, respondents must show that they are able to do some meaningful part of the assigned task.
- On average, respondents who receive partial credit on an item should have average literacy scores (based on the other items they completed) between the scores of respondents who got the item completely right and respondents who got the item wrong.

The analysis staff examined the 1992 rubrics for the six blocks of items that were reused in 2003 to see which items had score points that made substantive sense as partial credit. Six out of some 65 items were identified with score points that were substantively appropriate for partial credit:

- N100401 (block 10, question 4), score point 2
- N100601 (block 10, question 6), score point 2
- N110601 (block 11, question 6), score point 2
- N110901 (block 11, question 9), score point 2
- N120401 (block 12, question 4), score point 2
- N130301 (block 13, question 3), score points 2 and 3

The analysis staff eliminated N100601 from consideration because respondents who received a partial credit score had literacy levels almost identical to respondents who got the item entirely wrong. This indicated that the item would not be likely to scale as partial credit. For the remaining five items, respondents who received the potential partial credit score point had total block scores that fell below the total block score for respondents who answered the question correctly and above the total block score for respondents who answered the question incorrectly. Once the potential partial credit items were identified, the five items were recoded as partial credit.

14.3 SCALING METHODOLOGY

Following the procedures used in 1992, the dataset used for scaling included respondents who completed five or more tasks on each of the prose, document, and quantitative literacy scales. Before scaling began, the analysis staff scored the data following the 1992 guidelines:[2]

- The correct key(s) for the item were considered **Right**.
- Nonresponses *that were followed by valid responses to other items in the same block* were considered **Omitted**. Items scored as **Omitted** were treated as though they were **Wrong**.
- Nonresponses *that occurred **after** the last item in a block with a valid response* were considered **Not Reached**. Items scored as **Not Reached** were treated as though they had never been presented to the respondent. This was done so as not to underestimate the literacy ability of respondents who did not complete an entire block.

[2]Exceptions to the scoring rules were made for partial credit items (discussed in section 14.1.2.3), which were considered partially correct rather than right or wrong.

- Multiple responses were considered **Wrong**.
- "I Don't Know" responses were considered **Wrong.**
- All other responses were considered **Wrong**.

This section reviews the scaling model employed in the analyses of the NAAL data and describes the marginal maximum likelihood (MML) methodology used for proficiency estimation.

14.3.1 The Scaling Model

Two distinct scaling models, depending on item type and scoring procedure, were used in the analysis. Each model is based on item response theory (IRT). Each is a "latent variable" model, defined separately for each of the scales, which expresses respondents' tendencies to achieve certain scores (such as correct/incorrect) on the items contributing to a scale as a function of a parameter that is not directly observed, called proficiency (θ) on the scale.

The item response models used differ only in the form of the function $P_i(\theta)$. The two-parameter logistic (2PL) model (Hambleton, Swaminathan, and Rogers 1991), which was used for dichotomous items (that is, items that are scored either right or wrong), takes the form

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$$

Where $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ answers item i correctly and a_i and b_i are parameters characterizing item i in terms of its discrimination and difficulty. In the 1992 National Adult Literacy Survey, the three-parameter logistic (3PL) model was adopted to fit the dichotomous items. The 3PL model was not necessary for the NAAL 2003 calibration since no multiple choice items were included.

For the partial credit items, we used the Graded Response Logistic (GRL) model (Samejima, 1969, 1972). This model follows the 2PL model for the probability of a score of 1 (at least partially correct):

$$P_{i1}(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_{i1})}}$$

It also follows the 2PL model for the probability of a score of 2 (completely correct):

$$P_{i2}(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_{i2})}}$$

Linear transformation of the scales was used to link the NAAL scales to the 1992 NALS scales for gain purposes. The scale indeterminacy was resolved by setting an origin and unit size of theta to the reported scale means and standard deviations from 1992 NALS.

A basic assumption of item response theory is the conditional independence of the responses by an individual to a set of items, given the individual's θ score (a measure of proficiency). That is, item response probabilities depend only on the individual's θ and the specified item parameters, as opposed to depending on any demographic characteristics of examinees, the position of the item in the booklet, the content of items around an item of interest, or the assessment administration conditions. Conditional on θ, the probability of a correct response on one item is unrelated to the probability of a correct response on another. This allows the following formula for the joint probability of examinee *i*'s response pattern $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{in})'$, where z_{ij} is person *i*'s score on item *j*, across a set of *n* items for given ability θ and item parameters:

$$P(\mathbf{z}_i \mid \theta, \text{item parameters}) = \prod_{j=1}^{n} \prod_{k=0}^{m_j - 1} P_{jk}(\theta)^{v_{jk}}$$

where m_j is the number of score categories of item *j*, and $v_{jk} = \begin{cases} 1 \text{ if } z_{ij} = k \\ 0 \text{ otherwise} \end{cases}$.

After the hypothetical response pattern *z* is observed, the above function can be viewed as a likelihood function that is to be maximized with a given set of item parameters. These item parameters were treated as known for the subsequent calculations.

Another assumption underlying the model is unidimensionality—that is, performance on a set of items is accounted for by a single construct. Although this assumption may be too strong, the use of the model is motivated by the need to summarize overall performance parsimoniously within a single domain. Hence, item parameters were estimated for each scale separately.

Testing the assumptions of the item response theory model is a critical part of the data analyses. A number of checks were made to detect multidimensionality and certain condition dependencies. Differential item functioning (DIF) analyses were used to examine issues of dimensionality (see section 4.2 in chapter 4), and item fit was examined to flag responses with serious departures from the IRT model. The fit of the IRT models to the observed data was examined within each scale by comparing the

empirical item response functions (IRFs) with the theoretical curves. The theoretical curves were plots of the response functions based on the estimates of the item parameters. The empirical proportions were calculated from the posterior distributions of the theta for each student who received the item. For good fitting items, the empirical and theoretical curves were close together. Items for which this was not true indicated poor fit and were examined carefully. When warranted, remedial efforts, such as collapsing categories of polytomous item or removing items from the test, were made to mitigate the effects of such violations on inferences.

14.3.2 Linking to the 1992 NALS

As already noted, the prose, document, and quantitative literacy results from the NAAL are reported on scales that were established in the 1992 NALS; 65 (43 percent) of the tasks administered in the 2003 NAAL were originally administered in 1992. The linkage between the scales from the two studies is based on these tasks. In addition, 88 new tasks were developed for the NAAL and therefore a total of 153 tasks were administered in the 2003 assessment. However, out of the 88 new tasks, one task (C060401) was dropped prior to scaling on the basis of the examination of the item analysis statistics. A total of 152 tasks were retained for scaling.

14.3.3 Item Parameter Estimation

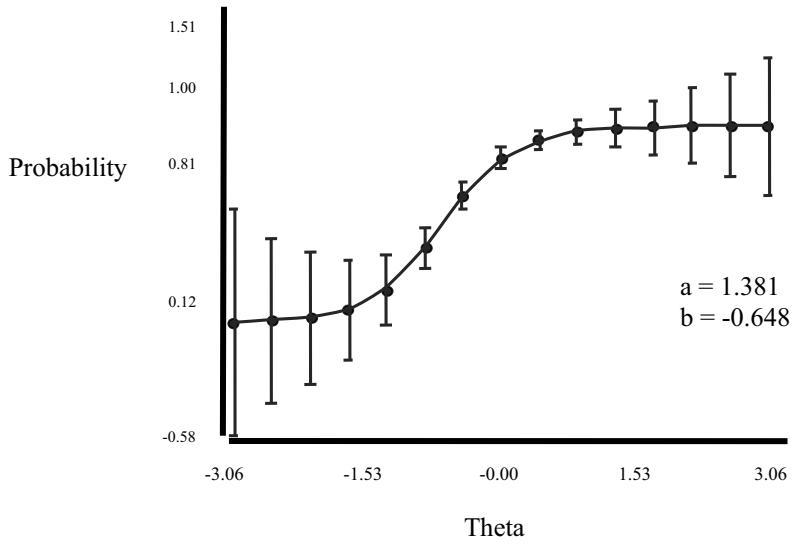
Identical item calibration procedures were carried out separately for each of the three literacy scales. Analysts used the IRT package of the AM software developed by Cohen et al. (2000). The two-parameter logistic item response theory model was fit to dichotomous items and the Graded Response Logistic item response theory model was fit to partial credit items. Preliminary sample weights were used during the calibration procedures.

After operational data were received by NAAL staff, all items were put into the scaling, with the exception of the one item that was identified as flawed on the basis of the item analysis (C060401). During the scaling, any items identified as problematic on the basis of the item analysis were closely watched to ensure that the scaling model was correctly fitting the data. At this stage, certain items scored as partial credit were not well fit by the IRT Graded Response Logistic model. These items were collapsed back to right/wrong items and the scaling of the full item set was repeated.

In addition, linking the 1992 and 2003 assessments required that items presented in both years could be fit by using the same IRT item parameters, based on data from both years. The fit of the model to any item that was given in both assessments was evaluated by checking the fit of the IRT model

estimated using both years of data to the data from each individual assessment. Any item showing lack of fit for either or both assessments was "split"; that is, it was treated as two distinct items: one for 1992 and another for 2003. Model fit was evaluated at the task level by inspecting residuals from fitted item response curves from AM. The item response curves were visually examined by comparing the empirical item response functions (IRFs) with the theoretical curves. An example of item response curve is presented in figure 14-1.

Figure 14-1. Example of item response curve (item N100201) from the NAAL: 2003



NOTE: Dots represent the 2003 NAAL data. They indicate estimated conditional probabilities obtained without assuming a specific model form. The curve indicates the estimated item response function (IRF) assuming a logistic model form. The bars around the dots indicate the standard errors around the dots.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

As a result of the waves of analysis, NAAL staff were able to successfully scale together the 1992 and the 2003 data. Only one item was dropped from scaling (C060401). Nine items in the 2003 dataset that were scored as partial credit were collapsed as a result of misfit when they were scaled with the graded response model. And only one item needed to be split (N130901 which was renamed CN130901 in the 2003 dataset where it was treated as a new item). Table 14-1 summarizes the treatment of the items (i.e., whether they were dropped, collapsed, or split). Estimated item parameters for each literacy scale are

presented in tables D-1 through D-3 in appendix D. As shown in appendix D, the slope or discrimination parameters (parameter a) range from 0.41 to 2.55 for the prose literacy scale, from 0.41 to 2.63 for the document literacy scale, and from 0.34 to 4.60 for the quantitative literacy scale. The difficulty parameters (parameter b) for dichotomous items range from -2.50 to 1.71 for the prose literacy scale, from -6.34 to 1.52 for the document literacy scale, and from -2.82 to 1.77 for the quantitative literacy scale. The ranges of the step parameters for polytomous items are from -1.92 to 1.62 for the prose literacy scale, from -2.03 to 1.16 for the document literacy scale, and from -1.76 to 0.74 for the quantitative literacy scale.

Table 14-1. NAAL items, by item treatment during scaling: 2003

Item treatment	Item
Dropped	C060401
Collapsed	C030301, C030707, C040401, C040503, C040701, C040801, C061001, C061101, C080503
Split	CN130901

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

14.4 DIRECT ESTIMATION OF SUBPOPULATION DISTRIBUTIONS OF PROFICIENCY USING DIRECT ESTIMATION WITH THE AM SOFTWARE

14.4.1 Background

As with most survey programs, NAAL faces competing demands. NAAL must contend with the paradoxical requirement that the test be both long enough to assess proficiency on a broad set of literacy skills and knowledge and short enough to ensure that the test can be completed in a reasonable length of time. The requirement for a short test comes from the very nature of the program: for examinees, NAAL is a low-stakes test, so they are unlikely to expend much effort. Individual examinees never receive any feedback about their individual performance; in fact, individual scores are never assigned at all. Therefore, the test must be short to avoid exceeding the effort that examinees are likely to expend.

NAAL fulfills both of these competing objectives by using an incomplete-block test design. Under this design, test items are arranged into large number of "blocks," and only a small number of the blocks appear in each test booklet. Data from the different booklets are tied to a common scale through statistical methods based on item response theory (IRT; Rasch, 1960; Lord, 1952; Birnbaum, 1968; Hambleton and Swaminathan, 1995). Under this design, each examinee completes a single, short booklet containing a subset of blocks. Across all booklets, the assessment includes many items, enough to cover

the full extent of the underlying trait being measured. Hence, NAAL appears short to the individual examinees—where it needs to be short—and long from the perspective of aggregate content coverage.

The use of IRT methods enables NAAL to calibrate items from different test booklets to reveal a common, partially observed latent trait, which we will call *proficiency*. Hence, although the measurement properties of the booklets vary both cross-sectionally and over time, the measurement models putatively yield access to a consistent underlying construct.

NAAL's incomplete-block design exacerbates a problem shared by all assessment programs that report aggregate proficiency statistics. Tests measure proficiency imperfectly, and the measurement error in individual scores can bias estimates of underlying proficiency, even when the individual-level instrument is very precise. For example, Mislevy, Beaton, Kaplan, and Sheehan (1992) estimate that the variability among individual scores from a 30-item test would have a reliability of 0.80 and would overestimate the population variance by 25 percent. Individual scores from an 85-item individual test with a reliability of 0.90 would still overestimate the variance by 11 percent. Similarly, the measurement error in these assessments would cause underestimates of correlation coefficients and bias other aggregate statistics such as the proportion of the population within specified score ranges. With the incomplete-block design, scores from the individual-level instruments tend to be quite imprecise, increasing the biases owing to measurement error. Further, the measurement properties typically differ across booklets, making the exact impact of the bias somewhat unpredictable.

In 1992, NALS used methods derived from Marginal Maximum Likelihood (MML) estimation to obtain unbiased estimates of the target statistics. Until recently, however, appropriate MML software was not widely available. Therefore, the contractor in charge of the 1992 NALS used the plausible values method developed for National Assessment of Educational Progress (NAEP; Mislevy 1984, 1985, 1991; Thomas, 1993) to allow secondary users to estimate statistics derived from individual data.

Plausible values are multiple imputations randomly drawn from a distribution derived from the MML parameter estimates for an extensive conditioning model (Allen, Carlson, and Zelenak, 1999). When analyzed as though they are test scores free of measurement error, plausible values yield good approximations of many aggregate statistics. However, it is important to understand that they remain approximations not estimated directly from the data.

Plausible values were introduced in the 1980s to provide secondary analysts as a way to analyze incomplete block-design assessment data using existing software and computers available at that time. Times have changed, however, in that most desktop computers provide more than enough power to

directly estimate statistical models that plausible values approximate. The following is a brief overview of the direct estimation of models based on MML using the AM software, along with a Taylor-series approximation to the standard error for these models.

14.4.2 Direct Estimation of MML Models

This section describes MML regression, as well as the numerical optimization method that AM uses to estimate the model. The model itself is not new, and the optimization method is well known; however, this section describes some of the computational advantages of this approach over the now-standard EM optimization. In addition, this section describes the application of Binder's (1983) implicit differentiation method for variance estimation in complex samples to the MML regression. The development of the variance estimator addresses the fact that virtually all large-scale assessments involve complex sample designs, generally invalidating standard variance estimates.

Based on section 14.3.1, the conditional probability of examinee *i*'s response pattern $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{in})'$, where z_{ij} is person *i*'s score on item *j*, across a set of *n* items for given ability θ and item parameters is:

$$P(\mathbf{z}_i \mid \theta, \text{item parameters}) = \prod_{j=1}^{n} \prod_{k=0}^{m_j - 1} P_{jk}(\theta)^{v_{jk}}$$

where m_j is the number of score categories of item *j*, and $v_{jk} = \begin{cases} 1 \text{ if } z_{ij} = k \\ 0 \text{ otherwise} \end{cases}$. Suppose that the θ is a random draw from a population distribution with probability density function $f(\theta \mid \text{population parameters})$, then following Bayes rule, the marginal likelihood function of the population parameters for person *i* is:

$$L_i(\text{population parameters} \mid \ _i, \text{item parameters}) \propto$$
$$\int p(\ _i \mid \theta, \text{item parameters}) f(\theta \mid \text{population parameters}) d\theta$$

In practice, a normal distribution with mean μ and standard deviation σ is often assumed for $f(\theta \mid \text{population parameters})$, hence, the marginal likelihood function becomes:

$$L_i(\mu, \sigma \mid \mathbf{z}_i, \text{item parameters}) \propto \int p(\mathbf{z}_i \mid \theta, \text{item parameters}) f(\theta \mid \mu, \sigma) d\theta \qquad (1)$$

Recall that in the regression case, $\mu_i = \mathbf{x}_i'\mathbf{b}$, where $\mathbf{x}_i$ is the design matrix formed by the independent variables of person *i*, and $\mathbf{b}$ is the regression parameters . In this case, the marginal likelihood function becomes:

$$L_i(\mathbf{b}, \sigma \mid \mathbf{z}_i, \mathbf{x}_i, \text{item parameters}) \propto \int p(\mathbf{z}_i \mid \theta, \text{item parameters}) f(\theta \mid \mathbf{x}_\text{i}; \mathbf{b}, \sigma) d\theta \qquad (1')$$

Note that likelihood function never actually requires a point estimate of proficiency (theta) for each individual. Rather, the method essentially "tries" all possible values, weighting each one by the probability that a random draw from a population with mean μ_i and variance σ^2 would yield it. In this way, the method estimates the distribution of proficiency in the population without ever estimating proficiency for each individual. This is why the method is called "marginal": it yields point estimates of the group or subgroup parameters without requiring point estimates for individual students by integrating the θ parameter out.

The estimation task for computing statistics based on θ involves specifying and maximizing the likelihood function across observations. For numerical reason, the log-likelihood function is often used in place of the likelihood function during the maximization procedure. We discuss the algorithm for doing so below, followed by a discussion of how we obtain estimates of the standard errors that are appropriate under a complex sample design.

14.4.2.1 Estimation

We estimate this model by using an algorithm that combines quasi-Newton optimization with numeric quadrature over about 30 equidistant points along the feasible range of θ. Although an EM algorithm is often used for models with the general form of equation 1, in this particular model it proves inefficient. The first term on the right-hand side of equation 1 is a product over many IRT link functions with fixed parameters. These are time-consuming computations. With fixed-distance quadrature points, this set of computations is required just once. Most EM algorithms would require recomputation at each iteration.

We use numeric quadrature to approximate the integral in equation 1, which is difficult to evaluate analytically. We identify a range within which all observations are virtually certain to fall and then select equally spaced points along this interval. For example, if theta is a standard normal (0, 1) variate, it makes sense to have the quadrature points range between about –4 and 4 (99.994% of cases will

fall in this range). Thus, letting $q = \{1,2,...Q\}$ for Q quadrature points (θ_q), we can rewrite the individual likelihood function (1') as

$$L_i(\mathbf{b},\sigma \mid \mathbf{z}_i,\mathbf{x}_i, \text{item parameters}) \approx \sum_{q=1}^{Q} p(\mathbf{z_i} \mid \theta_q, \text{item parameters}) f(\theta_q \mid \mathbf{x}_i;\mathbf{b},\sigma). \qquad \textbf{(2)}$$

The goal then is to find the values of the parameters ($\mathbf{b},\sigma$) that maximize the likelihood function. This is typically done by using iterative methods that try various values of the parameters, evaluate the likelihood function, then adjust the provisional parameter estimates to values that increase the function.

The method of Berndt, Hall, Hall, and Hausman (BHHH; 1974) was used to find values of $\mathbf{b}$ and σ that maximize the likelihood functions. (The algorithm employed here offers the option of taking a single steepest-descent step when a BHHH step fails to yield improvement in the likelihood function. Alternatively, users can select the slower steepest-ascent algorithm.) This method modifies Newton's method and has proven quite successful in a range of MML problems. Letting $\mathbf{\Gamma} = (\mathbf{b}',\sigma\ '$ represent a vector of the parameters, Newton's method uses the following iterations to update $\mathbf{\Gamma}$: $\mathbf{\Gamma_{t+1}} = \mathbf{\Gamma_t} - \mathbf{G_t^{-1}}\mathbf{g_t}$ for iteration $t+1$, based on values from iteration *t*. In this equation, $\mathbf{g}_t = \sum_i \mathbf{g}_{it}$ is the vector of first derivatives of the log-likelihood function with respect to the parameters from iteration *t*, and $\mathbf{G_t}$ is the matrix of second derivatives of the log-likelihood function with respect to the parameters from iteration *t*. Although Newton's method has many desirable properties in a range of maximum likelihood problems, it only works where $-\mathbf{G_t^{-1}}$ is positive definite. Unfortunately, this is not always the case, especially in early iterations. The BHHH method substitutes $\mathbf{H_t}$ for $-\mathbf{G_t^{-1}}$ where $\mathbf{H_t} = \left[\sum_i \mathbf{g}_{it}\mathbf{g}'_{it}\right]^{-1}$, thereby guaranteeing a positive definite matrix everywhere. At convergence, either matrix consistently estimates the inverse of the information matrix.

Using this method, it is necessary to specify the likelihood function and the first-order derivatives with respect to the vector $\mathbf{b}$ and the scalar σ (which are appended into a vector, **g**). These derivatives take a relatively simple form.

The first *k* elements of **g** represent $\frac{\partial \ln(L)}{\partial b_k}$. This portion of **g** is given by

$$\frac{\partial \ln(L)}{\partial \mathbf{b}} = \sum_i \mathbf{x}_i \left(\frac{1}{\sigma^3 L_i} \sum_q - \exp^{-\frac{(\mathbf{x}_i'\mathbf{b}-\theta_q)^2}{2\sigma^2}} \left(\mathbf{x}_i'\mathbf{b} - \theta_q\right) p_{iq} \right)$$

where L_i is the value of the likelihood function for examinee i and $p_{iq} = p(\mathbf{z}_i \mid \theta_q, \text{item parameters})$. The final element of ***g***, the first derivative with respect to sigma, is given by

$$\frac{\partial \ln(L)}{\partial \sigma} = \sum_i \left(\frac{1}{\sigma^4 L_i} \sum_q p_{iq} \left(\exp^{-\frac{(\mathbf{x}_i'\mathbf{b}-\theta_q)^2}{2\sigma^2}} \left(\mathbf{x}_i'\mathbf{b} - \theta_q\right)^2 - \exp^{-\frac{(\mathbf{x}_i'\mathbf{b}-\theta_q)^2}{2\sigma^2}} \sigma^2 \right) \right).$$

This approach works well even when starting with far less than optimal starting values. To improve performance, we begin with a rough approximation: we calculate an approximate score for each respondent as the weighted average of the values of the quadrature points, where the weights are given by p_{iq}. We then estimate an OLS regression against these "pseudo-scores" and take these as starting values for the **b** parameters. The regression root MSE, appropriately adjusted for the unreliability of the pseudo-scores, provides a starting value for σ.

14.4.3 Weighting and Variance Estimation in Complex Samples

The roots of the score functions constitute sets of *estimating functions* (Godambe, 1960, 1991). Godambe (1960) proves an optimal property of maximum likelihood by estimating functions in simple random samples—and Godambe and Thompson (1986) show that the standard π-weighted estimating function retains this optimal property in unequal probability samples. Thus, the log-likelihood function and its derivatives at each observation can be multiplied by the inverse probability of inclusion in the sample, yielding optimal estimating functions.

The score function provides an estimating equation by which consistent estimates of the finite population parameters may be obtained, even though their superpopulation counterparts may be fictional:

$$\hat{W}(\hat{\mathbf{\Gamma}}) = \mathbf{0} = \sum_{i=1}^{n} w_i \frac{\partial \log(L_i(\hat{\mathbf{\Gamma}} \mid \mathbf{z}_i, \mathbf{x}_i, \text{item parameters}))}{\partial \hat{\mathbf{\Gamma}}}$$

where w_i represents the sample weight, usually the inverse of an estimate of the probability of selection.

The lack of independence among the observations and the misspecification of the model render the traditional maximum likelihood variance estimator based on the inverse of the observed information

matrix useless. More appropriate is an approximate variance estimator based on Binder's (1983) method of implicit differentiation, which Godambe and Thompson (1986) suggest be applied to estimating functions in the presence of unequal weights. Binder begins with a Taylor series expansion of the estimation function around the true-value . A first-order expansion of a linear estimating equation $\hat{W}(\hat{\mathbf{\Gamma}}) = 0$ yields

$$0 \approx \hat{W}(\mathbf{\Gamma}) + \frac{\partial \hat{W}(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}) \quad .$$

Solving for $\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}$ yields

$$\hat{\mathbf{\Gamma}} - \mathbf{\Gamma} = -\left[\frac{\partial \hat{W}(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}}\right]^{-1} \hat{W}(\mathbf{\Gamma}) \quad ,$$

and taking variance of both sides,

$$Var(\hat{\mathbf{\Gamma}}) = \left[\frac{\partial W(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}}\right]^{-1} \Omega(\mathbf{\Gamma}) \left[\frac{\partial W(\mathbf{\Gamma})'}{\partial \mathbf{\Gamma}}\right]^{-1}$$

where $\Omega(\mathbf{\Gamma})$ is the variance of $W(\mathbf{\Gamma})$ across observations. Substituting expectations (estimates) in place of true values yields the proposed standard error estimator:

$$Var(\hat{\mathbf{\Gamma}}) = \hat{G}^{-1}(\hat{\mathbf{\Gamma}}) Var(\hat{g}(\hat{\mathbf{\Gamma}})) \hat{G}^{-1}(\hat{\mathbf{\Gamma}})$$

This is popularly known as a *sandwich estimator*. In this case, the outer terms are approximated by $\mathbf{H} = \left[\sum_i \mathbf{g}_i \mathbf{g}_i'\right]^{-1}$, where $\mathbf{g}_i$ is the vector of first derivatives of the log-likelihood function with respect to the converged parameter values, i.e. $\mathbf{g}_i = w_i \dfrac{\partial \log(L_i(\hat{\mathbf{\Gamma}} \mid \mathbf{z}_i, \mathbf{x}_i, \text{item parameters}))}{\partial \hat{\mathbf{\Gamma}}}$. The variance term in the center is the estimated variance/covariance matrix of the first derivatives. Note that this is simply the variance/ covariance matrix of a set of population totals (the summed first derivatives). In a stratified, clustered, unequally weighted sample, one can usually approximate this as the π estimator of the stratified, between primary sampling unit (PSU) variance.

Using the stratified, between- primary sampling unit (PSU) weighted estimator to estimate $\hat{\Omega}(\hat{\mathbf{\Gamma}})$, gives

$$\hat{\Omega}(\hat{\mathbf{\Gamma}}) = \sum_{h=1}^{H} (\frac{n_h}{n_h - 1}) \sum_{i=1}^{n_h} (\mathbf{g}_{hi} - \overline{\mathbf{g}}_h)(\mathbf{g}_{hi} - \overline{\mathbf{g}}_h)'$$

where $\mathbf{g}_{hi} = \sum_{k=1}^{m_{hi}} w_{hik} \frac{\partial \log(L_i(\mathbf{\Gamma} \mid \mathbf{z}_i, \mathbf{x}_i, \text{item parameters})}{\partial \mathbf{\Gamma}}$, and $\overline{\mathbf{g}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{g}_{hi}$, in which h indexes strata, i indexes primary sampling units, and k indexes individuals. In a simple random sample the PSUs are the sampled examinees.

14.5 LINKING THE 2003 SCALE TO THE 1992 SCALE

Initially, proficiency scores are estimated on the basis of a provisional scale with a mean of 0 and a standard deviation of 1. To be comparable to the scores from 1992, they need to be put on the same scale. This is accomplished through the use of linear transformation constants that match the mean and standard deviations of the 1992 sample based on the new item parameters.[3] The transformation that was applied is as follows:

$$\theta = \mathrm{A}\theta^* + \mathrm{B}$$

Where θ^* is the provisional scale from item calibration and θ is the reported 0 to 500 scale, and A and B are transformation constants. Table 14-2 presents the transformation constants (that is, the standard deviations and means) for the distributions of the three scales. These constants apply both to the 2003 data and to the 1992 data when the new item parameters are used.

Table 14-2. Transformation constants (standard deviations and means) using new item parameters, by literacy scale: 1992 and 2003

Literacy scale	A (SD)	B (Mean)
Prose	58.48056	280.705
Document	58.75546	274.8816
Quantitative	63.31159	280.4884
Composite	57.3496	280.6508

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey and 2003 National Assessment of Adult Literacy

[3]This is also based on the redefined sample of complete cases and rescored items from 1992 data.

14.6 MINIMUM SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In the NAAL reports, the sample sizes were not always large enough to permit accurate estimates of proficiency and/or background results for one or more categories of variables. For results to be reported for any subgroup, a minimum sample size of 45 was required. This number was arrived at by determining the sample size needed to detect an effect size of 0.5 with a probability of 0.8 or greater, using a design effect of 1.5. This design effect implies a sample design-based variance 1.5 times that of a simple random sample. The effect size of 0.5 pertains to the true difference in mean proficiency between the subgroup in question and the total population, divided by the standard deviation of proficiency in the total population. An effect size of 0.5 was chosen following Cohen (1988), who classifies effect size of this magnitude as "medium" as well as to be consistent with what was done in the 1992 survey.