CHAPTER 15

THE LITERACY OF ADULTS WITHOUT COGNITIVE DATA

Ying Jin and Stéphane Baldi, American Institutes for Research

15.1 INTRODUCTION

Missing data are always expected in any large-scale assessment. Sampled individuals may not respond to an assessment for many reasons. For example, they may not respond because they have a language barrier or a physical disability, or they may simply refuse. A number of alternative ways are available to deal with missing data. The least desirable way is simply to ignore the missing data. This practice assumes that the data are missing at random and that the remaining observed samples are representative of the target population. However, if the pattern of missing data is correlated to the outcome of the study, this practice would yield both biased and inaccurate estimates of proficiency distributions for some subpopulations and consequently for the total population as well. For further information on nonresponse bias analysis and different approaches to treating missing data, the reader can refer to section 4-4 of NCES Statistical Standards (NCES 2003-601) and Groves et al. (2001).

Experience with the 1992 NALS and other assessments, as well as evidence from the 2003 NAAL assessment, indicates that adults with lower levels of literacy are more likely than adults with higher proficiencies to refuse to respond to the assessment. Ignoring the pattern of missing data would have resulted in overestimating the literacy skills of adults in the United States. This chapter describes the methods adopted to deal with the problem of missing cognitive data.

15.1.1 Missing Background Data

The target sample for the 2003 NAAL assessment included 35,365 nationally representative housing units, of which 4,671 were vacant. Approximately 18 percent of the households that were occupied at the time of data collection refused to participate, and no detailed background information is available on this group. For the households that agreed to participate in the study, the interviewers began by using a series of screening questions to obtain an accurate count of the number of age-eligible persons in the household. Depending on the number of eligible persons in the household, one or two persons were selected to participate in the interview to complete the background questionnaire (BQ). Respondents who did not answer a sufficient number of background questions were considered incomplete cases. Cases that

were mostly incomplete could not be analyzed and were never incorporated into the database. Such cases were dealt with through weighting class and poststratification adjustments for instrument nonresponse (see chapter 12). For some background variables, such as education and country of birth, missing data for literacy-related reasons were imputed using logistic regression models (see chapter 12).

15.1.2 Missing Cognitive Data

The 19,714 adults ages 16 and older residing in households or prisons who agreed to respond to the assessment answered extensive background questions during the interview, as appropriate to the household or prison samples, about their age, country of birth, language(s) spoken or read, highest level of education completed, current educational aspirations, labor market status, current occupation and wages, voting behaviors, and reading habits. After answering the background questions, respondents were asked to complete the literacy tasks in the assessment booklet. Very easy tasks were placed first to encourage respondents to continue. (See chapter 2 for details on booklet and block design of the cognitive assessment.) Nevertheless, 1,155 (6 weighted percent) of these respondents did not complete any cognitive tasks on any of the three scales, and 256 (1 weighted percent) did not complete any cognitive tasks on at least one but not all scales. For individuals who refused to continue after answering the background questions, no information is available about their performance on the cognitive tasks. Completely omitting these individuals from the analyses would have resulted in overestimates of the literacy skills of the national population as a whole and particularly of certain subpopulations. Imputation procedures were applied to enable the estimation of their literacy proficiencies.

15.2 THE NORMAL TREATMENT OF MISSING COGNITIVE DATA

For the National Assessment of Adult Literacy, a distinction should be made between missing responses by design and missing responses by individual respondents who were presented with the cognitive questions. In population assessments, unlike in individual assessments, a matrix design for item sampling in which examinees respond to different subsets of cognitive questions is usually used to limit the burden on respondents. Because of the matrix sampling design of the NAAL assessment, each respondent received only a fraction (three-thirteenths) of the literacy tasks; therefore, most of the tasks were not presented to every respondent and could be considered missing. This type of missing data was intentional by design. Missing responses could also occur for the cognitive tasks that were presented to respondents.

15.2.1 Omitted, Not Reached

For the NAAL literacy tasks that were presented, the missing responses occurred in two distinct patterns: the respondent skipped over a question and responded to a subsequent question, or the respondent broke off the assessment and did not attempt to respond to any subsequent questions. The two types of nonresponse were called "omitted" and "not reached" tasks, categories that were based on the distinction between missing responses prior to the last question the respondent answered in each block and missing responses subsequent to the last observed response in each block.

- **Omitted.** In some cases, respondents skipped over a particular task but attempted or completed one or more tasks that followed. This kind of missing responses, called "omitted," occurred prior to the last observed response in each block and by definition could not be at the end of a block.

- **Not reached**. In other cases, respondents spent all their time responding to preceding tasks and did not reach tasks that appeared later in a block. This kind of missing responses was observed subsequent to the last question the respondent attempted in each block. Tasks that were not attempted were found consecutively at the end of the blocks and called "not reached**."**

In the omitted-response situation, there was a logical basis for assigning a wrong answer to a missing response. As in the 1992 NALS survey, in the 2003 NAAL assessment, omitted cognitive responses were treated as wrong answers, on the assumption that respondents decided to skip them intentionally because they found these tasks too difficult. The treatment of an omitted response as a wrong answer was a logical imputation that was based on the circumstances that surrounded the missing data.

In contrast, not-reached cognitive responses were not treated as wrong but were treated as if the questions had not been presented to the respondent. The assumption here was that respondents did not make a task-specific decision about whether to respond to tasks that were not reached. Because there was not a sufficient logical basis for assigning a wrong answer, these responses remained missing data. Moreover, it was unlikely that not-reached tasks occurred because of speededness—that is, because the respondent was not given enough time to answer them. The assessment booklet was not rushed and the NAAL was not a speeded test, but there were some practical time limits so that the interviewer would not have to spend an unreasonable number of hours collecting information.

15.2.2 Cognitive Data Missing by Design

The 2003 NAAL assessment had 153 individual literacy tasks. It was impractical to ask respondents to take the entire pool of the assessment questions, particularly for a low-stakes assessment such as the NAAL. Therefore, to allow maximum coverage of literacy materials and content while minimizing the time burden for any one respondent, the NAAL assessment used a matrix item sampling design (see chapter 7) in which individuals responded to different subsets of cognitive tasks. Because of the matrix design of the NAAL assessment, each respondent received only three-thirteenths of the literacy tasks. Therefore, for every respondent, most of the tasks were not presented. The tasks that were not presented could be considered missing, but unlike omitted or not-reached tasks, this type of missing data was intentional by design.

15.3 REASONS FOR MISSING DATA

As will be described in the next section, missing data were imputed on the basis of the reasons for nonresponse, in particular, whether or not the reason was literacy related. This section summarizes the reasons for missing data in the NAAL sample. Section 15.3.1 gives the distribution of the sample by presence of cognitive data and reasons for missing data. Section 15.3.2 provides support for the validity of reasons through a comparison of reasons for various demographic subgroups.

15.3.1 Disposition Codes and Literacy Skills

Table 15-1 shows the distribution of the final weighted NAAL sample into categories on the basis of the following criteria: presence of cognitive data, type of nonresponse, and response to the Adult Literacy Supplemental Assessment (ALSA). Literacy-related reasons for nonresponse consisted of language problems, mental disabilities (including mental retardation, learning disabilities, and mental/emotional conditions), and reading/writing barriers. Refusals, physical disabilities (including visual and hearing impairments), and other/unknown reasons were classified as not literacy related because the disposition code provided no direct evidence of low literacy skills.

The distribution in table 15-1 is provided for the total sample and separately for the household sample and the prison sample. Of the 19,714 cases in the total sample, 93 percent had complete cognitive data. Complete cognitive data were obtained for 93 percent of the household sample (total sample size 18,541) and for 96 percent of the prison sample (total sample size 1,173). For the remaining cases,

literacy scores were imputed if the sample person had partial cognitive data, the reason for incomplete cognitive data was not literacy related, or the sample person had completed the ALSA. There were 456 cases (3 percent) that did not meet these requirements and for whom scores were not imputed.

Table 15-1. Percent distribution of final weighted NAAL sample, by presence of cognitive data, reason for missing data, and sample type: 2003

			Incomplete cognitive data		No cognitive data		
Sample	Total	Complete cognitive data	Literacy related	Not literacy related[1]	Literacy related (no imputation)	Literacy related, ALSA[2] respondent	Not literacy related[1]
Total	100.0	92.6	0.2	0.7	2.9	0.1	3.6
Household	100.0	92.5	0.2	0.7	2.9	0.1	3.6
Prison	100.0	96.0	#	0.9	1.4	0.2	1.5

Rounds to zero.
[1] Disposition code provides no direct evidence of low literacy skills.
[2] Adult Literacy Supplemental Assessment.
NOTE: Incomplete cognitive data have at least one but not all scales. 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 disposition code assigned to the case was used to determine whether cognitive data were missing for literacy-related reasons. The same procedure was used for the household and prison samples. Table 15-2 shows the weighted distribution of the 19,714 NAAL sample cases by presence of cognitive data and the reason for incomplete data. Among cases with no cognitive data, the predominant reason for nonresponse was refusal (2.2 percent), followed by language problems (1.9 percent).

Table 15-2. Distribution of final weighted NAAL sample of 19,714, by presence of cognitive data and detailed reason for missing data: 2003

Presence of cognitive data and report reason (if no cognitive data)	Percent of adults	Presumed relation of reason to literacy skills
Total	100.0	†
Cognitive data	93.4	†
No cognitive data	6.6	†
Refused	2.2	Not literacy related[1]
Non-English language	1.9	Literacy related
Mental disability, including retardation, learning disability, and other mental/emotional condition	0.9	Literacy related
Other or unknown	0.7	Not literacy related[1]
Physical disability, including visual	0.7	Not literacy related[1]
Reading and/or writing difficulty	0.2	Literacy related

† Not applicable.
[1] Disposition code provides no direct evidence of low literacy skills.
NOTE: 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.

15.3.2 Household Sample Internal Evidence for the Validity of Reasons

The reasons provided for incomplete assessments were used to determine the treatment of cases with missing cognitive data. This section presents an evaluation of the validity of those reasons for the household sample.[1] An analysis was performed for cases with no cognitive data, using demographic information from the screener and the background questionnaire. The evaluation follows the analysis presented in chapter 8 of the Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey (Yamamoto 2001). Table 15-3 displays the distribution of the interviewer-coded reasons for not providing cognitive data, by age, race/ethnicity, language spoken when growing up, and educational attainment. Standard errors for table 15-3 were computed in WesVar, using replicate weights, and are provided in table 15-4.

[1] The analysis was restricted to the household sample because of small sample sizes and unstable estimates for the prison sample.

Table 15-3. Percentage of household study population by presence of cognitive data, reason for missing data, and population group: 2003

Population group	Total sample size	Percent with cognitive data present	Percent with no cognitive data present							
			Reasons related to literacy				Reasons unrelated to literacy			
			Total	Non-English language	Mental disability	Reading or writing difficulty	Total	Refused	Physical disability	Other, unknown, no answer
Total	18,500	93.4	3.0	1.9	0.9	0.2	3.6	2.2	0.7	0.7
Age (years)										
16–29	4,700	96.0	1.8	1.0	0.7	0.1	2.2	1.8	#	0.4
30–49	7,300	94.3	2.7	2.1	0.5	0.1	3.1	2.2	0.1	0.7
50–69	4,600	93.3	2.7	1.8	0.7	0.2	4.0	2.6	0.8	0.6
70+	2,000	85.2	7.5	3.3	3.4	0.9	7.3	2.4	3.5	1.4
Race/ethnicity										
Hispanic	3,200	91.0	3.8	2.3	0.8	0.7	5.2	2.7	0.8	1.8
Black	3,500	94.3	1.7	0.7	0.6	0.5	4.0	1.4	0.8	1.8
Asian	400	87.2	7.1	7.1	#	#	5.7	4.5	#	1.2
White/other[1]	11,400	94.0	2.9	1.8	1.0	0.1	3.1	2.2	0.6	0.3
Language spoken while growing up										
English	15,700	96.5	0.3	0.1	0.1	0.1	3.2	2.1	0.7	0.5
Other language	2,500	89.7	3.1	2.4	0.1	0.7	7.2	4.1	0.6	2.6
Missing	400	1.1	98.8	65.1	33.7	#	0.1	#	#	0.1
English spoken while growing up										
Hispanic	1,200	94.7	0.6	0.1	0.2	0.2	4.8	3.5	1.0	0.3
Black	3,400	95.5	0.5	#	#	0.5	4.0	1.4	0.8	1.8
Asian	200	94.3	2.7	2.7	#	#	3.0	3.0	#	#
White/other[1]	10,900	96.8	0.2	#	0.1	0.1	3.0	2.1	0.7	0.3

See notes at end of table.

Table 15-3. Percentage of household study population by presence of cognitive data, reason for missing data, and population group: 2003—Continued

Population group	Total sample size	Percent with cognitive data present	Percent with no cognitive data present							
			Reasons related to literacy				Reasons unrelated to literacy			
			Total	Non-English language	Mental disability	Reading or writing difficulty	Total	Refused	Physical disability	Other, unknown, no answer
English not spoken while growing up										
Hispanic	1,900	92.7	1.5	0.5	0.1	1.0	5.8	2.2	0.7	2.9
Black	100	94.0	#	#	#	#	6.0	2.0	2.0	2.0
Asian	200	79.6	11.9	11.9	#	#	8.6	6.1	#	2.4
White/other[1]	300	86.5	1.9	1.6	#	0.3	11.7	9.3	0.5	1.9
Education[2]										
Less than high school	4,600	87.7	7.3	4.7	1.7	0.9	5.0	1.7	1.3	2.0
High school or GED	4,700	93.3	2.5	1.7	0.7	0.1	4.2	2.8	0.8	0.7
More than high school	9,200	95.8	1.6	0.9	0.7	#	2.6	2.1	0.3	0.2

#Rounds to zero.
[1]Other includes non-Hispanic American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, and multiple races.
[2]Includes values of educational attainment imputed during the weighting process for weighting purposes.
NOTE: 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 15-4. Standard errors for table 15-3

Population group	Standard error of percent with cognitive data present	Standard error of percent with no cognitive data present							
		Reasons related to literacy				Reasons unrelated to literacy			
		Total	Non-English language	Mental disability	Reading or writing difficulty	Total	Refused	Physical disability	Other unknown, no answer
Total	0.44	0.26	0.26	0.09	0.04	0.37	0.35	0.08	0.08
Age (years)									
16–29	0.57	0.27	0.21	0.16	0.05	0.49	0.48	0.03	0.09
30–49	0.45	0.33	0.31	0.09	0.04	0.41	0.39	0.06	0.11
50–69	0.71	0.41	0.39	0.19	0.07	0.57	0.49	0.17	0.13
70+	1.33	0.87	0.70	0.49	0.25	0.78	0.49	0.45	0.31
Race/ethnicity									
Hispanic	1.05	0.83	0.79	0.25	0.20	0.77	0.70	0.26	0.25
Black	0.57	0.36	0.29	0.12	0.20	0.45	0.22	0.22	0.36
Asian	2.75	2.03	2.03	#	#	1.43	1.29	#	0.63
White/other[1]	0.49	0.28	0.27	0.10	0.03	0.40	0.36	0.10	0.06
Language spoken while growing up									
English	0.35	0.06	0.05	0.04	0.04	0.33	0.31	0.09	0.07
Languages other than English	1.44	0.71	0.64	0.04	0.20	1.08	1.03	0.21	0.38
Missing	0.65	0.65	3.66	3.54	#	0.07	#	#	0.07
English spoken while growing up									
Hispanic	1.34	0.26	0.06	0.20	0.17	1.26	1.17	0.43	0.13
Black	0.49	0.20	0.04	0.02	0.21	0.44	0.21	0.22	0.36
Asian	2.21	1.29	1.29	#	#	1.38	1.38	#	#
White/other[1]	0.38	0.06	0.03	0.05	0.03	0.36	0.32	0.10	0.06

See notes at end of table.

Table 15-4. Standard errors for table 15-3—Continued

Population group	Standard error of percent with cognitive data present	Standard error of percent with no cognitive data present							
		Reasons related to literacy				Reasons unrelated to literacy			
		Total	Non-English language	Mental disability	Reading or writing difficulty	Total	Refused	Physical disability	Other unknown, no answer
English not spoken while growing up									
Hispanic	0.88	0.35	0.16	0.06	0.31	0.73	0.62	0.30	0.39
Black	3.03	#	#	#	#	3.03	2.06	1.67	1.55
Asian	4.64	3.70	3.70	#	#	2.27	2.00	#	1.18
White/other[1]	2.89	0.89	0.86	#	0.24	2.90	2.86	0.48	0.94
Education[2]									
Less than high school	0.78	0.72	0.66	0.29	0.19	0.51	0.37	0.24	0.26
High school or GED	0.64	0.38	0.34	0.13	0.04	0.55	0.47	0.15	0.13
More than high school	0.45	0.18	0.16	0.11	0.01	0.39	0.39	0.07	0.04

#Rounds to zero.
[1]Other includes non-Hispanic American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, and multiple races.
[2]Includes values of educational attainment imputed during the weighting process for weighting purposes.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Age and Reasons. For adults ages 70 and over, physical disabilities were the predominant reason for not providing any cognitive data. The prevalence of missing cognitive data related to physical disabilities increased with age, from less than 0.1 percent for sample persons ages 16 to 29 to 3.5 percent for those ages 70 and older. Mental disabilities, such as dementia, may also be associated with aging. As is evident in table 15-3, adults ages 70 and older with no cognitive data were more likely than younger adults to give mental disabilities and reading or writing difficulties as a reason for nonresponse. Overall, adults ages 70 and over were about as likely to give a literacy-related reason as a non-literacy-related reason, whereas the other age groups were more likely to provide reasons unrelated to literacy. The same pattern was evident in the 1992 NALS. Although the trend seen in the 1992 NALS of increasing refusals with increasing age was evident in the current study for the lower three age groups, it did not hold true for the group ages 70 and over.

Race/Ethnicity and Reasons. Four racial/ethnic groups were evaluated: Hispanic, Black, Asian, and White/other. The "other" group included non-Hispanic American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, and multiple races. Non-Hispanic White and "other" were combined because of the small sample size for the "other" category. Speaking a non-English language was provided as a reason for having no cognitive data for 7.1 percent of the Asian population, compared with 2.3 percent of the Hispanic population, 1.8 percent of White/other populations, and 0.7 percent of the Black population. Asian adults were also the only racial/ethnic group more likely to be assigned a literacy-related than a non-literacy-related disposition code when cognitive data were missing. Differences among racial/ethnic groups may be attributable to differences in languages spoken while growing up.

Native Language and Reasons. For sample persons with no cognitive data, language problems were given as a reason for nonresponse for approximately one-third of adults who did not speak English while growing up, consistent with the 1992 NALS analysis. In contrast, less than 0.1 percent of adults who spoke English while growing up had language problems as a reason for not providing cognitive data.

Native Language, Race/Ethnicity, and Reasons. When the population was restricted to adults who spoke English while growing up, only minor differences were seen in the reasons for missing cognitive data among the Hispanic, Black, and White/other racial/ethnic groups. The percentage of language problems remained higher among the Asian population than among the other racial/ethnic groups, but the sample size for this group was small. For adults who did not speak English while growing up, the differences were greater, with 0.5 percent of the Hispanic population, 11.9 percent of the Asian population, and 1.6 percent of the White/other population having no cognitive data because of problems with the English language. Among the small Black population who did not speak English while growing up, no sample persons lacked cognitive data as a result of language problems.

Education and Reasons. The proportion of sample persons who were missing cognitive data because of literacy-related reasons was greater among those with less than a high school education than among those with higher educational attainment. Reading/writing difficulties contributed to the lack of cognitive data for less than 0.1 percent of the population with a high school degree or higher.

Disposition codes were related to the age, race, language, and education of the sampled persons. In addition, the relationships were consistent with those of the 1992 NALS evaluation as noted above. The notable exception was the greater percentage of Hispanic adults with cognitive data in the 2003 NAAL, which was a result of the implementation of the ALSA. Overall, the analysis lends support to the validity of reasons provided for missing cognitive data and to the potential bias that could be incurred if missing data were ignored.

15.4 POPULATION ESTIMATES FOR ADULTS WITH UNMEASURABLE LITERACY SKILLS

In the 2003 NAAL assessment, among the 19,714 persons who agreed to respond to the assessment, 1,155 (6 percent) of these respondents did not complete any cognitive tasks on any scale, and 256 (1 percent) did not complete any cognitive tasks on at least one scale. For such nonrespondents, no information was available about their performance on the literacy scale that they did not complete.

When missing data patterns are related to the outcome of a study, the missing responses, if ignored, will bias the results unless some adjustment can be made to counter the bias. Evidence from the 2003 NAAL assessment indicated that response rate were different among subpopulations, and adults with lower levels of literacy were more likely than adults with higher proficiencies either to decline to respond to the assessment at all or to begin the assessment but not complete it. Excluding these individuals from the analyses would have resulted in overestimates of the literacy proficiencies of the national population as a whole and particularly of certain subpopulations.

15.4.1 Using "Reasons" to Improve Treatment of Missing Cognitive Data

As in the 1992 NALS, nonrespondents in the 2003 NAAL were classified into two separate groups: people who did not answer any cognitive questions for literacy-related reasons and people who did not respond for non-literacy-related reasons, based on the adults' self-reported reasons for nonresponse that Westat collected in the field. Literacy-related reasons included reading/writing barrier, language problem, and mental disability. Non-literacy-related reasons included physical disability (respondents were offered assistance with writing if necessary), refusals, unavailability for the field

period, and other. Responses to the background variables indicated that those who did not respond to the cognitive items for literacy-related reasons were disproportionately likely to be foreign born, to have less than a high school education, to be Hispanic or Asian/Pacific Islander, and to be age 65 or older. These variables are known to relate to English language proficiency and cognitive skills. Combined with other background information, there was strong evidence to support the notion that nonresponse to the cognitive items was not a random occurrence.

To enable the estimation of the literacy proficiencies of the nonrespondents, AIR applied both logical and regression-based imputation procedures to the missing responses on the basis of the type of reasons for nonresponse (see Little and Rubin 1987 for various imputation methods). The data concerning reasons for missing cognitive data provided the basis for making logical imputations of what the missing answers would have been had the respondent completed the assessment booklet. In the regression-based imputation strategy, the NAAL background data for the nonrespondents were used to fit the logistic regression models. In each of the logistic regression models, the dependent variables were the items selected to be imputed. The variables used to perform nonresponse and poststratification adjustments were used as predictors. Two different sets of predictor variables were used for respondents from the household sample and from the prison sample because the background questionnaires were somewhat different. Tables 15-5 and 15-6 describe the predictor variables used in the logistic regression models in the Household Study and the Prison Study, respectively.

SAS procedures for logistic regression were used to conduct the logistic regression analyses for each item to be imputed. The estimated regression coefficients were used to predict missing values of the dependent variables. For each nonrespondent, the probability of answering the item correctly was computed and then compared with a randomly generated number between 0 and 1. If the probability of getting a correct answer was greater than the random number, the imputed value for the item was 1 (correct). Otherwise, it was 0 (wrong).

All respondents who answered at least one item on each scale (prose, document, and quantitative) were included in the main NAAL reporting sample only on the basis of their performance on the items they answered, including respondents who answered only core items and were put in the ALSA. For the remaining respondents, the specific imputation procedures and decisions were as follows[2:]

(A) For ALSA respondents who had missing responses to any core items:

[2] The decision to impute only the easiest item on each scale (and no more) was made so as to have the same number of respondents on every scale while limiting estimation error due to imputed data.

- Impute all missing core items as wrong on all three scales. It was logical to assign wrong answers to ALSA nonrespondents because they were the least literate adults and most of the ALSA respondents answered the NAAL questions incorrectly (which was how they were classified into ALSA).

Table 15-5. Variables used as predictors in the logistic regression models for imputation in the NAAL household study, by value label: 2003

Variable (NAME)	Value label
Census Region (REGION)	1: Northeast
	2: Midwest
	3: South
	4: West
Metropolitan Statistical Area Status (MSAFLG)	0: NonMSA
	1: MSA
Age From Background Data (AGECAT)	1: 16–29
	2: 30–49
	3: 50–69
	4: 70+
Gender From Background Data (GENDER_R)	1: Male
	2: Female
Highest Education Level From Background Data (EDUCCAT)	1: Less than high school
	2: High school diploma or equivalent
	3: More than high school
Country of Birth (BORNUSA)	0: Born elsewhere
	1: Born in the USA
Race-Ethnicity From Background Data (BQ_RACETH_RAKE)	1: Hispanic
	2: Non-Hispanic Black only
	3: Other

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

Table 15-6. Variables used as predictors in the logistic regression models for imputation in the NAAL prison study, by value label: 2003

Variable (NAME)	Value label
Prison Security Level (SECURITY_R)	1: Supermax, Max/close/high
	2: Medium
	3: Minimum/low, Admin, Other
Region/Prison Type (REGTYPE)	1: Northeast
	2: Midwest
	3: South
	4: West
	5: Federal
Inmate Age Category (AGECAT_R)	1: 16–29
	2: 30–49
	3: 50+
Inmate Gender (GENDER_R)	1: Male
	2: Female
Inmate Highest Education Level (EDUCCAT)	1: Less than high school
	2: High school or higher
Inmate Country of Birth (BORNUSA)	0: Born elsewhere
	1: Born in the USA
Inmate Race-Ethnicity (BQ_RACETH_RAKE)	1: Hispanic
	2: Non-Hispanic Black only
	3: Other
Inmate Marital Status (MARITAL)	1: Never Married
	2: Other

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

(B) For NAAL respondents (not put into ALSA) who did not attempt any item on at least one scale for non-literacy-related reasons, impute one item on the scale(s) for which they had no responses (with one exception for the document scale as described below):

- Impute easiest core item on prose (CC004) and quantitative (CC005) scales via the logistic regression models described above.
- Impute easiest health item on document scale (CC007) via the logistic regression model described above (even if another document item had been answered).

(C) For NAAL respondents (not put into ALSA) who did not attempt any item on at least one scale but not all scales for literacy-related reasons:

- Impute one item as wrong on the scale(s) for which they had no responses as discussed in (B) above.

(D) For NAAL respondents who did not attempt any item on any scale for literacy-related reasons:

- Do not impute. Leave as missing.

15.4.2 Maintaining Comparability of Population Coverage Between 1992 and 2003 Assessments

The NAAL 2003 reporting sample included respondents from the household survey and the prison survey, including those who were identified as least literate and were administered the ALSA assessment. It was possible and reasonable to include ALSA respondents in the NAAL reporting sample because ALSA respondents met the minimum requirement of answering at least one question on each scale because they all took the seven core NAAL questions before they were classified into ALSA. Another reason to include ALSA in the NAAL reporting sample was to facilitate the identification of a comparable population from the 1992 NALS in order to report trend between the two years.

In the 1992 NALS, approximately 5 percent (weighted) of respondents were unable to complete the simplest literacy tasks in the assessment and were classified in the lowest level of literacy. No separate assessment comparable to ALSA was developed in 1992 to measure the literacy skills of those lowest literacy adults. If the ALSA respondents were not to be included in the NAAL reporting sample, a population comparable to the 2003 ALSA would have to be identified from the 1992 NALS data to ensure that trend could be maintained between the two assessments. AIR explored various ways to identify such a comparable population in the 1992 NALS data and concluded that including ALSA respondents in the 2003 NAAL reporting sample was the most appropriate solution to the problem of identifying the matching 1992 reporting sample for trend study.

As indicated in the previous section, a different imputation strategy was applied in the 1992 NALS from that used in the 2003 NAAL. To report trend to 1992, AIR applied the same imputation procedures used in the 2003 NAAL to the 1992 NALS data and identified a population in the 1992 NALS comparable to the 2003 assessment (i.e., a population that included individuals who answered at least one item on any scale).

The section below describes the specific imputation decisions for the 1992 NALS:

(A) For respondents who did not answer any questions on any scale for non-literacy-related reasons:

- Impute the easiest core item on prose (NC00301), quantitative (NC00501), and document (SCOR100) scales via logistic regression models. The 1992 NALS background data were used to fit the logistic regression models in which the dependent variables were the three imputed items, respectively. The variables used for nonresponse and poststratification adjustments were used as predictors. Tables

15-6 and 15-7 list the predictor variables used in the logistic regression models for the 1992 NALS household study and prison study, respectively. The variables were listed separately for respondents from the household sample and from the prison sample because the background questionnaires were somewhat different.

(B) For respondents who did not answer any questions on any scale for literacy-related reasons:

- Do not impute. Leave as missing.

(C) For respondents who did not answer any questions on at least one scale but not all scales:

- If reasons for nonresponse were not literacy related: Impute the easiest core item on the scale(s) for which they had no responses via logistic regression models as discussed in (A) above.
- If reasons for nonresponse were literacy related: Impute the easiest core item on the scale(s) for which they had no response as wrong.

Table 15-7. Variables used as predictors in the logistic regression models for imputation in the NALS household study, by value label: 1992

Variable (NAME)	Value label
Census Region (REGION)	1: Northeast
	2: Midwest
	3: South
	4: West
Age From Background Data (AGE)	1: 16–29
	2: 30–49
	3: 50–69
	4: 70+
Gender From Background Data (GENDER_R)	1: Male
	2: Female
Highest Education Level From Background Data (EDUCATION)	1: Less than high school
	2: High school diploma or equivalent
	3: More than high school
Race-Ethnicity From Background Data (RACE-ETHNICITY)	1: Hispanic
	2: Non-Hispanic Black only
	3: Other

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

Table 15-8. Variables used as predictors in the logistic regression models for imputation in the NALS prison study, by value label: 1992

Variable (NAME)	Value label
Prison Security Level (FACTYPE)	1: Maximum
	2: Medium
	3: Minimum
	4: Other
	5: Medical Center
Census Region (REGION)	1: Northeast
	2: Midwest
	3: South
	4: West
Inmate Age Category (AGE)	1: <30
	2: Other
Inmate Gender (GENDER)	1: Male
	2: Female
Inmate Highest Education Level (EDUCATION)	1: Less than high school
	2: High school or equivalent
	3: More than high school
Inmate Race-Ethnicity (RACE-ETHNICITY)	1: Hispanic
	2: Non-Hispanic Black only
	3: Other

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

CHAPTER 16

VARIABLE CONSTRUCTION AND FILE DEVELOPMENT

Ying Jin, American Institutes for Research

This chapter discusses the variable construction and file development procedures for the National Assessment of Adult Literacy (NAAL). The discussion begins with the construction and the contents of the NAAL public use data for the Household Study and the Prison Study as well as the NAAL item parameter files, followed by a discussion on the construction procedures for the derived variables (see appendix A for details). Also in this chapter are instructions on how to analyze NAAL data by using AM software (see appendix B for details) and how to use electronic codebooks (see appendix C for details).

16.1 FILE CONSTRUCTION PROCEDURES AND FILE CONTENTS

The NAAL study includes three components: (1) a national household sample; (2) a state household sample from six states, called the State Assessment of Adult Literacy (SAAL); and (3) a national sample of prison inmates. The NAAL and SAAL household samples were combined to allow improved precision of statistical estimates. Data from respondents were collected through a Background Questionnaire (BQ) and a literacy assessment. Other types of data were also collected and derived during the NAAL sampling, data collection, and weighting processes. On the basis of the National Center for Education Statistics (NCES) guidelines for data release, two separate public use data files containing the data above were created: one for the Household Study and one for the Prison Study. The household and prison samples were combined for the national estimates.

Using the guidelines to limit the risk of data disclosure provided by NCES standard 4-2, Maintaining Confidentiality, staff conducted the NAAL disclosure control analysis to reduce the risk of data disclosure and to protect the confidentiality of all individuals responding to the NAAL (see chapter 10 for details). Public data files for both the Household Study and the Prison Study were constructed on the basis of the recommendations from this statistical disclosure control analysis and therefore could be released to the public. Restricted-use data files that contained more complete and detailed data, including data that were suppressed in the public data files because of disclosure risk reasons, were available at the NCES website for licensed users only. For procedures to obtain an NCES license for restricted use data, the reader can go to http://nces.ed.gov/StatProg/confid.asp.

16.1.1 Household Study

The public use data file for the Household Study contains the following four types of variables:

- ID: Randomly assigned unique case identifier.
- Assessment items: Cognitive tasks from the NAAL literacy assessment on each of the three literacy scales (prose, document, and quantitative). Each task measuring health literacy was also classified as a prose, document, or quantitative task and was included on one of those three scales. Each assessment item variable contains the scored data from respondents (right, wrong, omitted, not-reached, and missing).
- Background variables: Variables containing data collected from the NAAL Household Background Questionnaire (BQ). The NAAL household BQ covered a variety of areas including demographic information, language background, educational background and experiences, health, labor force participation, and so on, as discussed in section 2.2.
- Assessment design variables: Variables related to complex sampling design, including sampling weights, variance stratum variable, and cluster variable. These variables should be used appropriately for the analysis of the NAAL data to obtain the most accurate estimates.

16.1.2 Prison Study

Like the public use data file for the Household Study, the data file for the Prison Study contains the four types of variables discussed above. However, the background variables contain data from the NAAL Prison Background Questionnaire, which collected demographic data on inmates and contextual data on their experiences in and/or prior to admission to prison that were related to literacy (see section 2.3 for details).

16.1.3 NAAL Item Parameter Files

Two item parameter files, one for the prose, document, and quantitative literacy scales and one for the health literacy scale, were created for the analysis of NAAL data using the AM software (see section 16.3). Such item parameter files are also called AM dictionary files and have an extension of .dct in their file names. The item parameter file for the health literacy scale was created separately from the other three scales because the health items were also classified as prose, document, or quantitative items and confounding those items into one file would cause problems in AM. The item parameter files contain the mean, the standard deviation, and the item parameters from scaling procedures (see chapter 14) for

each of the literacy scales. Appendix B describes how to use these files when using AM to analyze the NAAL data.

16.2 CONSTRUCTION OF DERIVED VARIABLES

The NAAL collected a large amount of information on many background variables. Some variables, such as gender, can be used directly for reporting purposes. Many variables, however, need to be derived from the raw data directly collected in the assessment for reporting purposes. For example, the NAAL collected information on respondents' date of birth. However, to report the literacy of adults in each age group of interest, a derived age variable needed to be constructed on the basis of the date of an interview and the respondent's date of birth. Derived variables were also constructed to make the trend variables comparable between the 2003 NAAL and the 1992 National Adult Literacy Survey (NALS) for trend analysis.

Appendix A contains the construction procedures of the derived variables that appear in each of the household and prison data files.

16.3 ANALYZING 2003 NAAL DATA USING AM SOFTWARE

The NAAL used a complex assessment design that allowed maximum coverage of the broad domain of literacy while minimizing the time burden on any one participant (see chapter 14). Under this design, the NAAL administered only a fraction of the assessment items on each scale to each participant. Although individual participants were required to take only a small portion of the entire pool of assessment questions, the aggregate results across the entire assessment allow broad reporting of literacy for the targeted population. However, because participants did not receive enough literacy tasks to provide reliable information about individual performance, traditional test scores for individual participants will result in misleading estimates of population characteristics and therefore are not appropriate to use for estimates of population statistics. Rather, statistical procedures based on the method of marginal maximum likelihood (MML) need to be used to provide consistent estimates of population statistics from data collected under such design. The usual statistical software packages such as SAS, SPSS or STATA cannot implement MML procedures; therefore, special analysis tools are needed. The AM software, which is available at http://am.air.org/, can be used to implement these procedures. The NAAL data that can be used in AM are available from NCES.

Appendix B has instructions to help first-time AM users analyze the NAAL data. The software also has an interactive help system that explains both how to use the software and the statistical procedures themselves.

16.4 USING THE ELECTRONIC CODEBOOK

The NAAL public use data for the Household Study and the Prison Study are available from NCES and can be accessed with an electronic codebook produced by NCES staff. The electronic codebook provides the option of producing SPSS, SAS, and STATA control statements that can be used to create SPSS, SAS, and STATA data files. These control statements also include statements for the following features:

VARIABLE DEFINITION	The field names are listed in the order in which they appear on the file. The electronic codebook will produce control statements in SPSS, SAS, or STATA with column positions and input formats.
VARIABLE LABEL	This is a 40-character text description for each field.
VALUE LABEL	All numeric fields with discrete (or categorical) values have 20-character text descriptors for each value within the variable's range.

In the electronic codebook, sections with variables for the household sample are arranged first, followed by variables for the prison sample. The data for the household and prison samples are stored in separate files. The electronic codebook also contains unweighted descriptive statistics of all variables. The electronic codebook is available in a Windows version. See appendix C for a more comprehensive user's manual for the electronic codebook.

CHAPTER 17

THE NAAL HEALTH LITERACY COMPONENT

Ying Jin and Eric Dunleavy, American Institutes for Research

Understanding the health literacy of America's adults is important because so many aspects of health care, from finding health information to maintaining health, depend on understanding written information. To determine the health literacy of the nation's adults and directly compare it with the measures of the general literacy of the population (i.e., prose, document, and quantitative), the 2003 National Assessment of Adult Literacy (NAAL) included a new component—a health literacy component. This chapter describes the definition of health literacy, construction of the health literacy scale, and reporting of the results on the health literacy scale.

17.1 DEFINING HEALTH LITERACY AND DETERMINING THE CONTENT OF THE HEALTH LITERACY COMPONENT

The NAAL health literacy scale and health literacy tasks were guided by the definition of health literacy used by the Institute of Medicine and *Healthy People 2010* (a set of national disease prevention and health promotion objectives led by the U.S. Department of Health and Human Services). This definition states that health literacy is:

> The degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions. (U.S. Department of Health and Human Services 2000 and Institute of Medicine 2004).

The content of the health literacy component was determined by the U.S. Department of Health and Human Services (HHS) in accordance with the public health priorities represented in *Healthy People 2010*, the disease prevention and health promotion agenda for the nation, and in consultation with HHS staff and external health literacy experts. The health literacy component of the NAAL offers a vehicle by which HHS can obtain a baseline measurement of the U.S. population's health literacy skills and project a target for improvement by the end of the decade (see section 2.4.3 in chapter 2).

17.2 IDENTIFYING HEALTH LITERACY TASKS

Twenty-eight tasks were included in the health literacy scale. These health literacy tasks represent a range of literacy activities that adults are likely to face in their daily lives. In the development of the health literacy tasks, the Office of Disease Prevention and Health Promotion (ODPHP) within the U.S.

Department of Health and Human Services suggested materials and questions based on input from other HHS agencies and stakeholders and experts, and on information from federal health materials and other health-related assessments.

17.2.1 Overlap of the health literacy tasks with the tasks on the prose, document and quantitative scales

The health literacy stimulus materials and the 28 tasks were designed to elicit respondents' skills for locating and understanding health-related information and services and to represent the three general literacy scales—prose, document, and quantitative—developed to report NAAL results. That is to say, the health literacy tasks were developed to fit into the NAAL's prose, document, or quantitative scales but were distinguished from the other tasks on those scales by their health content. The health literacy scale consisted of 12 prose, 12 document, and 4 quantitative NAAL tasks. As described in chapter 2, prose, document, and quantitative literacy were defined as follows:

- *Prose literacy*. The knowledge and skills needed to perform prose tasks (i.e., to search, comprehend, and use information from continuous texts). Prose examples include editorials, news stories, brochures, and instructional materials. Prose texts can be further broken down as expository, narrative, procedural, or persuasive.

- *Document literacy.* The knowledge and skills needed to perform document tasks (i.e., to search, comprehend, and use information from noncontinuous texts in various formats). Document examples include job applications, payroll forms, transportation schedules, maps, tables, and drug and food labels.

- *Quantitative literacy*. The knowledge and skills required to perform quantitative tasks (i.e., to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials). Examples include balancing a checkbook, figuring out a tip, completing an order form, and determining the amount of interest on a loan from an advertisement.

17.2.2 Framework of the health literacy scale

Tasks used to measure health literacy were organized around three domains of health and health care information and services: *clinical, prevention,* and *navigation of the health care system*.

The materials were selected to be representative of real-world health-related information, including insurance information, medicine directions, and preventive care information.

Of the 28 health literacy tasks, 3 represented the *clinical* domain, 14 represented the *prevention* domain, and 11 items represented the *navigation of the health care system* domain. The domains are defined in the following way:

- The *clinical* domain encompasses those activities associated with the health care provider-patient interaction, clinical encounters, diagnosis and treatment of illness, and medication. Tasks from the clinical domain are filling out a patient information form for an office visit, understanding dosing instructions for medication, and following a health care provider's recommendation for a diagnostic test.

- The *prevention* domain encompasses those activities associated with maintaining and improving health, preventing disease, intervening early in emerging health problems, and engaging in self-care and self-management of illness. Examples are following guidelines for age-appropriate preventive health services, identifying signs and symptoms of health problems that should be addressed with a health professional, and understanding how eating and exercise habits decrease risks for developing serious illness.

- The *navigation of the health care system* domain encompasses those activities related to understanding how the health care system works and individual rights and responsibilities. Examples are understanding what a health insurance plan will and will not pay for, determining eligibility for public insurance or assistance programs, and being able to give informed consent for a health care service (U.S. Department of Health and Human Services 2003, p. 37).

The NAAL health literacy scale did not include tasks that did not fit the definitions of prose, document, or quantitative literacy even if they were consistent with the definition of health literacy used by *Healthy People 2010*. For example, none of the NAAL health tasks required knowledge of specialized health terminology. The assessment also did not measure the ability to obtain information from non-print sources, although questions about the use of all sources of health information—both written and oral—were included on the background questionnaire.

17.3 SCALING THE HEALTH ITEMS

Since health literacy is distinguished from the prose, document and quantitative literacy with its unique health content, a separate health literacy scale was established to represent this unique latent trait. Similar item calibration procedures used for the other three literacy scales (see chapter 14) were carried out for the health items.

17.3.1 The scaling model

As discussed in chapter 14, each respondent did not answer all of the NAAL items under the NAAL assessment design and a simple average percent correct would not allow for appropriate reporting of population characteristics. Item response theory (IRT) methods, which models the probability of answering a question correctly as a mathematical function of proficiency or skill, were used to scale the health items as well as the prose, document, and quantitative items. The IRT scaling procedures provide a common scale on which performance on some latent trait can be compared across groups.

IRT models assume that an examinee's performance on each item reflects characteristics of the item and characteristics of the examinee. All models assume that all items on a scale measure a common latent ability or proficiency dimension (e.g., health literacy) and that the probability of a correct response on an item is uncorrelated with the probability of a correct response on another item given fixed values of the latent trait. Items are measured in terms of their difficulty as well as their ability to discriminate among examinees of varying ability.

The same IRT models used for scaling the prose, document, and quantitative items were used for the health literacy scale: the two-parameter logistic (2PL) model and the Graded Response Logistic (GRL) model, depending on the item types (i.e., dichotomous or polytomous). Each model is a "latent variable" model, 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. Details of the two scaling models were described in chapter 14.

The health literacy scale was established for the first time in the 2003 NAAL. Therefore, unlike the prose, document, and quantitative scales, there was no previous health assessment scale to link to. The reporting metric for the health literacy scale was set to have a mean of 245 and a standard deviation of 55, in order to resemble the reporting metrics of the prose, document and quantitative scales.

17.3.2 Item parameter estimation

As mentioned above, similar item calibration procedures for each of the prose, document and quantitative scales were performed for the health literacy scale. The IRT package of the AM software developed by Cohen et al. (2000) was used to fit the IRT models to the 2003 NAAL health assessment data. The two-parameter logistic item response theory model was adopted for dichotomous items and the Graded Response Logistic item response theory model was used for partial credit items. Model fit was

evaluated at the item 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. As in the scaling for the other NAAL literacy scales, preliminary sample weights were used during the calibration procedures.

One difference in item calibration procedures between the health literacy scale and the other NAAL scales is that, unlike the concurrent calibration procedures for the prose, document and quantitative scales that used the assessment data from both 1992 and 2003 (see chapter 14), the calibration of the health items was conducted only on the 2003 data. This was because the health scale was new to the 2003 assessment and there was no such previous scale in the 1992 assessment.

Estimated item parameters for the health literacy scale are presented in table D-4 in appendix D. As shown in appendix D, the slope or discrimination parameters (parameter a) range from 0.34 to 2.57 for the health literacy scale. The difficulty parameters (parameter b) for dichotomous items range from -7.11 to 1.52. The step parameters for polytomous items range from -2.07 to 1.64. Some items, e.g., CC001 in table D-2 and CC007 in table D-4, had extremely low values of difficulty parameters as such items were designed to be extremely easy to discriminate adults at the very low end of the literacy scale.

17.4 REPORTING RESULTS ON THE HEALTH LITERACY SCALE

In addition to average literacy scores, the NAAL results were also reported as the percentage of adults in the pre-defined performance levels. Performance levels are used to identify and characterize the relative strengths and weaknesses of adults falling within various ranges of literacy ability. Describing the adult population according to such levels allows analysts, policy-makers, and others to examine and discuss the typical performance and capabilities of specified groups within the adult population.[1]

17.4.1 Establishing health literacy performance levels

In response to the request of the Department of Education, the National Research Council's Board on Testing and Assessment (BOTA) Committee on Performance Levels for Adult Literacy recommended a new set of performance levels for the prose, document, and quantitative scales for the 2003 NAAL assessment, instead of using the same reporting levels used for the 1992 National Adult Literacy Survey. Hauser et al. (2005) described in detail the procedures followed by the BOTA Committee to determine the NAAL performance levels.

[1] For more information on NAAL performance levels, see White and Dillow (2005).

The new set of performance levels for each of the prose, document and quantitative scales are: *Below Basic, Basic, Intermediate,* and *Proficient*. Table 17-1 summarizes the knowledge, skills, and capabilities that adults needed to demonstrate to be classified into one of the four levels on the prose, document, and quantitative scales. However, The BOTA Committee was not asked to recommend performance levels for the health literacy scale. Because every health literacy task was included on the prose, document, or quantitative scale, it was assigned a performance level (*Below Basic, Basic, Intermediate,* and *Proficient*) corresponding to their position on one of those scales.

Table 17-1. NAAL literacy levels, by definition and key abilities associated with each level: 2003

Level and definition	Key abilities associated with level
Below Basic indicates no more than the most simple and concrete literacy skills. Score ranges for *Below Basic*: Prose: less than or equal to 209 Document: less than or equal to 204 Quantitative: less than or equal to 234	Adults at the *Below Basic* level range from being nonliterate in English to having the abilities listed below: ▪ locating easily identifiable information in short, commonplace **prose** texts ▪ locating easily identifiable information and following written instructions in simple **documents** (e.g., charts or forms) ▪ locating numbers and using them to perform simple **quantitative** operations (primarily addition) when the mathematical information is very concrete and familiar
Basic indicates skills necessary to perform simple and everyday literacy activities. Score ranges for *Basic*: Prose: 210–264 Document: 205–249 Quantitative: 235–289	▪ reading and understanding information in short, commonplace **prose** texts ▪ reading and understanding information in simple **documents** ▪ locating easily identifiable **quantitative** information and using it to solve simple, one-step problems when the arithmetic operation is specified or easily inferred
Intermediate indicates skills necessary to perform moderately challenging literacy activities. Score ranges for Intermediate: Prose: 265–339 Document: 250–334 Quantitative: 290–349	▪ reading and understanding moderately dense, less commonplace **prose** texts as well as summarizing, making simple inferences, determining cause and effect, and recognizing the author's purpose ▪ locating information in dense, complex **documents** and making simple inferences about the information ▪ locating less familiar **quantitative** information and using it to solve problems when the arithmetic operation is not specified or easily inferred
Proficient indicates skills necessary to per-form more complex and challenging literacy activities. Score ranges for Proficient: Prose: 340–500 Document: 335–500 Quantitative: 350–500	▪ reading lengthy, complex, abstract **prose** texts as well as synthesizing information and making complex inferences ▪ integrating, synthesizing, and analyzing multiple pieces of information located in complex **documents** ▪ locating more abstract **quantitative** information and using it to solve multi-step problems when the arithmetic operations are not easily inferred and the problems are more complex

NOTE: Although the literacy levels share common names with the National Assessment of Educational Progress (NAEP) levels, they do not correspond to the NAEP levels.
SOURCE: Hauser, R.M, Edley, C.F. Jr., Koenig, J.A., and Elliott, S.W. (Eds.). (2005). Measuring Literacy: Performance Levels for Adults, Interim Report. Washington, DC: National Academies Press; White, S. and Dillow, S. (2005). Key Concepts and Features of the 2003 National Assessment of Adult Literacy (NCES 2006-471). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

To determine the performance levels and cut scores on the health literacy scale, "item mapping" procedure was used. With this procedure, the health items were "placed" at a specific point on the health literacy scale based on their level of difficulty. The level of difficulty for each item was estimated as the proficiency level (i.e., score) at which the probability of answering an item correctly was 67 percent. The response probability (RP) of 67 percent – abbreviated as RP67– was the convention used by the BOTA Committee to map items on the prose, document, and quantitative scales (see Hauser et al. 2005). Using a response probability convention, it is possible to use the IRT model to place the items at some specific level on the scale, which in turn allows one to make statements or predictions about the likelihood that a person who scores at the level will answer the questions correctly.

The scale score associated with a 67 percent probability of a correct response was calculated for each health item based on the item parameters obtained from the item calibration process. The items were then rank-ordered based on the RP67 scale scores and mapped to the health literacy scale. Cut-points for performance levels on the health scale were established so that each item was classified into the same level on the health scale as on the respective prose, document, or quantitative scale with which the item was associated. Table 17-2 presents the item map for the heath literacy items. For dichotomous items, the RP67 scores are associated with a 67 percent probability of answering the item correctly. For partial credit items, the RP67 scores take into account the different score points and are associated with a 67 percent probability of a fully correct response and a partially correct response, respectively.

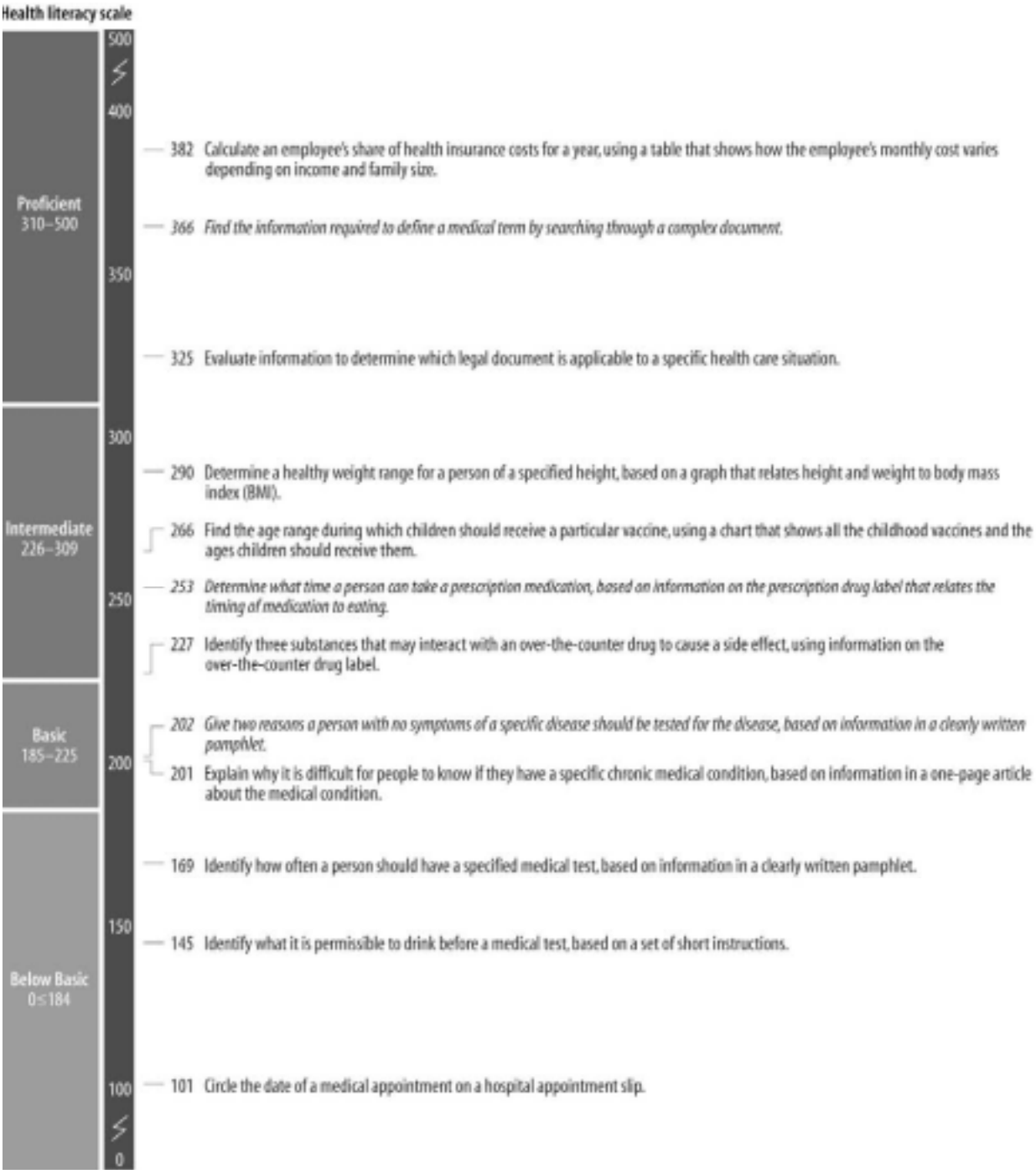
For further illustration, figure 17-1 shows the positions of selected health tasks on the health literacy scale. The position of a question on the scale represents the average scale score attained by adults who had a 67 percent probability of successfully answering the question. Next to each task is an indication of whether it was classified into the *Below Basic*, *Basic*, *Intermediate*, or *Proficient* category on the other scale on which the task was included (prose, document, or quantitative).

Table 17-2. Item map for the NAAL health literacy items: 2003

Item ID	Performance level on other scale	Item type	Score category	RP67 Scale score
CC007	Below Basic Document	Dichotomous	1	-58.7
CC002	Below Basic Document	Dichotomous	1	101.3
C070101	Below Basic Prose	Dichotomous	1	145.3
C071001	Below Basic Document	Partial credit	1	154.6
C020901	Below Basic Prose	Partial credit	1	157.9
C030201	Below Basic Prose	Dichotomous	1	168.8
C030101	Below Basic Prose	Partial credit	1	169.7
C070901	Below Basic Prose	Dichotomous	1	177.5
C050801	Below Basic Prose	Dichotomous	1	178.7
C020901	Basic Prose	Partial credit	2	186.3
C040501	Basic Document	Partial credit	1	191.3
N110101	Basic Prose	Dichotomous	1	201.0
C040504	Basic Document	Dichotomous	1	201.5
C030101	Basic Prose	Partial credit	2	202.3
C030301	Basic Prose	Dichotomous	1	202.4
C021001	Basic Document	Dichotomous	1	203.5
C071101	Basic Prose	Dichotomous	1	224.4
C080101	Basic Quantitative	Partial credit	1	225.0
C080201	Intermediate Document	Dichotomous	1	227.1
C060501	Intermediate Document	Dichotomous	1	238.8
C040503	Intermediate Document	Dichotomous	1	250.4
C080101	Intermediate Quantitative	Partial credit	2	253.1
C040502	Intermediate Document	Dichotomous	1	258.1
C060601	Intermediate Document	Dichotomous	1	265.6
C051101	Intermediate Prose	Dichotomous	1	275.9
C050901	Intermediate Prose	Dichotomous	1	277.2
N110201	Intermediate Quantitative	Dichotomous	1	280.8
C021101	Intermediate Document	Dichotomous	1	290.0
C040601	Intermediate Quantitative	Dichotomous	1	290.3
C040501	Intermediate Document	Partial credit	2	296.5
C051001	Proficient Prose	Dichotomous	1	324.7
C071101	Proficient Prose	Partial credit	2	366.1
C040801	Proficient Quantitative	Dichotomous	1	382.1

NOTE: For dichotomous items, the score category of 1 corresponds to the score point for correct responses. For partial credit items, the score category of 1 corresponds to the score point for partially correct responses, and the score category of 2 corresponds to the score point for fully correct responses. The RP67 scale score for item CC007 is extremely low (negative) because that item was designed to be extremely easy to discriminate adults at the very low end of the literacy scale.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

Figure 17-1. Difficulty of selected NAAL health literacy tasks: 2003



NOTE: The position of a question on the scale represents the average scale score attained by adults who had a 67 percent probability of successfully answering the question. Only selected questions are presented. Scale score ranges for performance levels are referenced on the figure. Regular type denotes a dichotomous item. Italic type denotes a partial credit item.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 National Assessment of Adult Literacy.

For example, as shown in figure 17-1, a task that requires a respondent to "evaluate information to determine which legal document is applicable to a specific health care situation" mapped to 325 on the health scale. That task was also included on the prose scale and, on the basis of the performance levels set by the BOTA Committee on Performance Levels for Adult Literacy, it mapped to the *Proficient* category on the prose scale. On the health scale, the cut-point between *Proficient* and *Intermediate* was set so that this task would fall into the *Proficient* category. That is, a health literacy task that was mapped to the *Proficient* level on the prose scale was also mapped to the *Proficient* level on the health scale.

Similarly, as shown in figure 17-1, a task that requires a respondent to "determine a healthy weight range for a person of a specified height, based on a graph that relates height and weight to body mass index (BMI)" mapped to 290 on the health scale. This task was also included on the document scale, where it was classified as *Intermediate*. The cut-points for the health scale were set so that the task would also map to the *Intermediate* level on the health scale.

As shown in figure 17-1, *Below Basic* health tasks required locating straightforward pieces of information in short simple texts or documents. Adults with literacy near the top of the *Below Basic* health literacy level were more likely to accomplish these tasks than adults who placed at the bottom of the *Below Basic* level.

Health tasks that mapped to the *Basic* level generally required finding information in texts and documents that were somewhat longer than those in the *Below Basic* level, and the information to be found was usually more complex. For example, a task that mapped to the *Basic* level required giving two reasons a person with no symptoms of a specific disease should be tested for the disease by using information in a pamphlet, while a task that mapped to the *Below Basic* level required finding one piece of information–the date–on a medical appointment slip that was shorter and simpler than the text in the *Basic* task. Health tasks that mapped to the *Intermediate* level went beyond simply searching texts and documents to find information. Most health tasks that mapped to the *Intermediate* level required adults to interpret or apply information that was presented in complex graphs, tables, or other health-related texts or documents. Health tasks that mapped to the *Proficient* level required drawing abstract inferences, comparing or contrasting multiple pieces of information within complex texts or documents, or applying abstract or complicated information from texts or documents.

Based on the procedures discussed above, the score ranges for each of the health performance levels were as follows: *Below Basic* 0-184, *Basic* 185-225, *Intermediate* 226-309, and *Proficient* 310-500.

17.4.2 Relationship of health scale to prose, document, and quantitative scales

Given the overlap of health items with prose, document, and quantitative items, it was expected that the correlations between the health literacy scale and the prose, document, and quantitative scales would be quite high. For reference purposes, table 17-3 presents these correlations.

Table 17-3. Correlations among NAAL literacy scales, by literacy scale: 2003

Literacy scale	Health	Prose	Document	Quantitative
Health	1.000			
Prose	0.940	1.000		
Document	0.998	0.857	1.000	
Quantitative	0.891	0.875	0.894	1.000

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

REFERENCES

Allen, N.L., Carlson, J.E., and Zelenak, C.A. (1999). *The NAEP 1996 Technical Report* (NCES 1999-452). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Bader, L.A. (2002). *Bader Reading and Language Inventory.* Upper Saddle River, MN: Prentice Hall.

Baker, D.W., Gazmararian, J.A.,Williams, M.V., Scott, T., Parker, R.M., Green, D., Ren, J., and Peel, J. (2004). Health Literacy and Use of Outpatient Physician Services by Medicare Managed Care Enrollees. *Journal of General Internal Medicine, 19*(3): 215–220.

Barton, D., and Hamilton, M. (1998). *Local Literacies: Reading and Writing in One Community.* London: Routledge.

Berndt, E.K., Hall, B.H., Hall, R.E., and Hausman, J.A. (1974). Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement, 4*: 653–665.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimates for Complex Surveys. *International Statistical Review, 51*(3): 279–292.

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

Burke, J., Mohadjer, L., Green, J., Waksburg, J., Kirsh, I., and Kolstad, A. (1994). Composite Estimation in National and State Surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 873–878.

Burns, P., and Roe, B.D. (2002). *Informal Reading Inventory: Preprimer to Twelfth Grade* (sixth edition). Boston: Houghton Mifflin.

Campbell, A., Kirsch, I.S., and Kolstad, A. (1992). *Assessing Literacy: The Framework for the National Adult Literacy Survey (*NCES 92113). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Cochran, W.G. (1977). *Sampling Techniques* (third edition). New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Jiang, T., Gaidurguis, A., and Hollender, D. (2000). AM (Software for marginal maximum likelihood and other statistical analyses of data from complex samples), Washington, DC: American Institutes for Research. See website at http://am.air.org.

CTB/McGraw-Hill. (2000). *Fox in a Box: An Adventure in Literacy*. Monterey, CA: Author.

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. *Review of Educational Research*, *49*(2): 222–251.

Ericsson, K., and Simon, H. (1980). Verbal Reports as Data. *Psychological Review, 87*(3): 215–251.

Fuchs, L., Fuchs, D., Hosp, M., and Jenkins, J. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading*, *5*(3): 239–256.

Godambe, V.P. (1960). An Optimum Property of Regular Maximum Likelihood Estimation. *Annals of Mathematical Statistics, 31*(4): 1208–1212.

Godambe, V.P. (1991). *Estimating Functions*. Oxford, U.K.: Clarendon Press.

Godambe, V.P. and Kale, B.K. (1991). Estimating functions: An overview. In *Estimating Functions* (V.P. Godambe, Ed.), pp. 3-20, Oxford, U.K.: Clarendon Press.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. *International Statistical Review, 54*(2): 127–138.

Good, R.H., Kaminski, R.A., and Dill, S. (2002). DIBELS Oral Reading Fluency. In R.H. Good and R.A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (sixth edition). Eugene, OR: Institute for the Development of Educational Achievement. From http://dibels.uoregon.edu.

Gordon, M.M., Hampson, R., Capell, H.A., and Madhok, R. (2002). Illiteracy in Rheumatoid Arthritis Patients as Determined by the Rapid Estimate of Adult Literacy in Medicine (REALM) Score. *Rheumatology (Oxford),41*(7): 750–754.

Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (2001). *Survey Nonresponse*. New York: John Wiley & Sons, Inc.

Halliday, M.A.K. (2002–). *The Collected Works of M. A. K . Halliday*. Ed. by Jonathan Webster. London: Continuum.

Halliday, M.A.K., and Hasan, R. (1985). *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press.

Hambleton, R.K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hambleton, R.K., and Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hamilton, M. (2000). Exploring Literacy as Social Practice through Media Photographs. In D. Barton, M. Hamilton, and R. Ivanic (Eds.), *Situated Literacies* (pp 16–34). London: Routledge.

Hauser, R.M, Edley, C.F. Jr., Koenig, J.A., and Elliott, S.W. (Eds.). (2005). Measuring Literacy: Performance Levels for Adults, Interim Report. Washington, DC: National Academies Press

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Boston: Heinle & Heinle Publishers.

Hill, C., and Parry, K. (1992). The Test at the Gate: Models of Literacy in Reading Assessment. *TESOL Quarterly, 26*(3): 433–461.

Holland, P.W., and Thayer, D.T. (1988). Differential Item Performance and Mantel-Haenszel. In H. Wainer, and H. Braun (Eds.). *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

Institute of Medicine (2004). *Health literacy: A Prescription to End Confusion.* Washington, DC: Institute of Medicine, Board on Neuroscience and Behavioral Health, Committee on Health Literacy.

Johns, J. (2001). *Basic Reading Inventory: Pre-Primer through Grade 12 & Early Literacy Assessments* (eighth edition). New York: Kendall Hunt Publishing Company.

Kirsch, I., Yamamoto, K., Norris, N., Rock, D., Jungeblut, A., O'Reilly, P., Berlin, M., Mohadjer, L., Waksberg, J., Goksel, H., Burke, J., Rieger, S., Green, J., Klein, M., Campbell, A., Jenkins, L., Kolstad, A., Mosenthal, P., and Baldi, S. (2001). *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey* (NCES 2001-457). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Kucera, H., and Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Leslie, L. and Caldwell, J. (2001). *Qualitative Reading Inventory-3*. New York: Addison Wesley Longman, Inc.

Lindau, S.T., Tomori, C., McCarville, M.A., and Bennett, C.L. (2001). Improving Rates of Cervical Cancer Screening and Pap Smear Follow-Up for Low-Income Women With Limited Health Literacy. *Cancer Investigation, 19*(3): 316–323.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons.

Lord, F.M. (1952). *A Theory of Test Scores.* Iowa City, IA: The Psychometric Society No. 7.

Marcus, M.P., Santorini, B., and Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2): 313–330.

Mislevy, R.J. (1984). Estimating Latent Distributions. *Psychometrika*, *49*(3): 359–381.

Mislevy, R.J. (1985). Estimation of Latent Group Effects. *Journal of the American Statistical Association*, *80*(392): 993–997.

Mislevy, R.J. (1991). Randomization-Based Inference About Latent Variables from Complex Samples. *Psychometrika*, *56*(2): 177–196.

Mislevy, R.J., Beaton, A.E., Kaplan, B., and Sheehan, K.M. (1992). Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29*(2): 133–161.

Muraki, E., and Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [computer program]. Chicago, IL: Scientific Software.

Oh, H.L., and Scheuren, F.J. (1987). Modified Raking Ratio Estimation. *Survey Methodology, 13*(2): 209-219.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Denmarks Paedagogiske Institut.

Reder, S and Edmonston, B. (2000). *Demographic Changes and Literacy Development in a Decade* (NCES 2000-09). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

Sabatini, J.P., and Venezky, R.L. (2000b). *Study of Adult Reading Acquisition (SARA): Test Development and Technical Data.* Philadelphia: University of Pennsylvania, National Center on Adult Literacy.

Sabatini, J.P., Venezky, R.L., Kharik, P., and Jain, R. (2000a). *Cognitive Assessments for Low Literate Adults: An Analytic Review and New Framework* (Technical Report TR00-01). Philadelphia: University of Pennsylvania, National Center on Adult Literacy.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph,* No. 18.

Shanker, J.L., and Ekwall, E.E. (2000). *Ekwall/Shanker Reading Inventory*. Boston: Allyn & Bacon.

Sheehan-Holt, J. K., and Smith, M. C. (2000). Does Basic Skills Education Affect Adults' Literacy Proficiencies and Reading Practices? *Reading Research Quarterly, 35*(2): 226–243.

Sherman, R., Condelli, L., and Koloski, J. (1999). *Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders* (NCES 1998–17). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

Shinn, M. R. (Ed.). (1989). *Curriculum-Based Measurement: Assessing Special Children*. New York: Guildford Press.

Silvaroli, N., and Wheelock, W.H. (2001). *Classroom Reading Inventory, Form C* (tenth edition). Boston: McGraw Hill.

Street, B. (1998). New Literacies in Theory and Practice: What are the Implications for Language in Education? *Linguistics and Education, 10*(1): 1–24.

Street, B. (2001). Introduction. In B. Street (Ed.), *Literacy and Development: Ethnographic Perspectives* (pp. 21–26). London: Routledge.

Swanson, H.L., Trainin, G., Necoechea, D.M., and Hammill, D.D. (2003). Rapid Naming, Phonological Awareness, and Reading: A Meta-Analysis of the Correlation Evidence. *Review of Educational Research, 73*(2), 407-440.

Thomas, N. (1993). Asymptotic Corrections for Multivariate Posterior Moments with Factored Likelihood Functions. *Journal of Computational and Graphical Statistics*, *2*(3): 309–322.

Texas Education Agency. (2001-2002). *Texas Primary Reading Inventory (TPRI)*. Austin, TX: Texas Education Agency.

U.S. Department of Education, National Center for Education Statistics. (2003). *NCES Statistical Standards* (NCES 2003-601). Washington, DC: Author.

U.S. Department of Health and Human Services (HHS). (2000). *Healthy People 2010: Understanding and Improving Health.* Washington, DC: U.S. Government Printing Office. Retrieved May 23, 2007 from http://www.healthypeople.gov/Publications/.

U.S. Department of Health and Human Services (HHS). (2003). *Communicating Health: Priorities and Strategies for Progress*. Washington, DC. Retrieved June 15, 2006 from http://odphp.osophs.dhhs.gov/projects/HealthComm/.

U.S. Office of Management and Budget. (1997, October 30). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. *The Federal Register*. Washington, DC: Government Printing Office.

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

White, S., and McCloskey, M. (2006 forthcoming). *Framework for the 2003 National Assessment of Adult Literacy* (NCES 2006-473). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Widdowson, H.G. (1983). *Learning Purpose and Language Use*. Oxford: Oxford University Press.

Widdowson, H.G. (1990). *Aspects of Language Teaching*. Oxford: Oxford University Press.

Williams, M.V., Davis, T., Parker, R.M., and Weiss, B.D. (2002). The Role of Health Literacy in Patient-Physician Communication. *Family Medicine, 34*(5), 383–389.

Woods, M.L., and Moe, A.J. (2003). *Analytical Reading Inventory* (eighth edition). New Jersey: Merrill/Prentice Hall.

Yamamoto, K. (2001). Estimating Literacy Proficiencies With and Without Cognitive Data. In I. Kirsch, K. Yamamoto, N. Norris, D. Rock, A. Jungeblut, P. O'Reilly, M. Berlin, L. Mohadjer, J. Waksberg, H. Goksel, J. Burke, S. Rieger, J. Green, M. Klein, A. Campbell, L. Jenkins, A. Kolstad, P. Mosenthal, and S. Baldi (Eds.) *Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey* (NCES 2001–457) (pp.142-164). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Appendix A

DERIVED VARIABLES

Appendix A: Derived Variables

This appendix describes how the derived variables that appear in each of the NAAL household and prison public use data files were constructed. References such as A-3 in the description for DAGE and K-3 in the description for DRACE for the household sample correspond to the question ID on the 2003 NAAL household background questionnaire (BQ). Similarly, references such as A-3 in the description for DAGEC and J-1 in the description for DRACE for the prison sample correspond to the question ID on the 2003 NAAL prison background questionnaire. The 2003 NAAL household and prison background questionnaires can be found at http://nces.ed.gov/NAAL/index.asp?file=DesignDevelop/SInstruments/BackQuestion.asp&PageID=116.

1. Derived variables for the NAAL 2003 household sample

DAGE (Age: 6 Categories)

Recoded from a continuous age variable derived from date of interview and the BQ response for date of birth, A-3. Missing values were filled in by using the age information from the Screener.

DAGEC (Age: 4 Categories)

Recoded from a continuous age variable derived from date of interview and the BQ response for date of birth, A-3. Missing values were filled in by using the age information from the Screener.

DRACE (Race/Ethnicity)

Derived from K-3 and K-5. If K-3 = 1, then DRACE = 3. Otherwise, if K-5A = 1 and K-5B-E = missing, then DRACE = 1; if K-5B = 1 and (K-5A and K-5C-E = missing), then DRACE = 2. Otherwise, DRACE = 4.

Note:

1. White, Black and Other categories include no Hispanics.

2. When the Ethnicity question K-3 = 7/8, the ethnicity indicated in the Screener was substituted.

   Likewise, if K-5A-E = 7/8, the race indicated in the Screener was substituted. Note that the data on race and ethnicity from the Screener may have not been reported by the respondent. That is, the person answering the Screener could have answered the race and ethnicity questions for another person in the household who was selected as a respondent. In these cases, it is not completely certain that the respondent would agree with the race/ethnicity designations recorded during the screener.

DARRIVE (Age Arrived In The U.S.)

Derived from A-1 and A-1A. If A-1 = “UNITED STATES (50 STATES OR DC),” then DARRIVE = 1. Otherwise, recode continuous responses for A-1A to categorical. If A-1A = missing, then DARRIVE = missing.

DENGAGE (Age Learned To Speak English)

Derived from A-8. If A-8 = missing, then DENGAGE = 1. These are people who speak only English. Otherwise, if A-8 = 1/2, then DENGAGE = 2; if A-8 = 3/4/5, then DENGAGE = 3; if A-8 = 95, then DENGAGE = 4; if A-8 = 97/98, then DENGAGE = missing.

DLIVEUS (Years Living In The U.S.)

Derived from A-1 and A-2. If A-1 = “UNITED STATES (50 STATES OR DC),” then DLIVEUS = 1. Otherwise, if A-2 = 1, then DLIVEUS = 2; if A-2 = 2/3/4/5/6/7/8, then DLIVEUS = 3; if A-2 = missing, then DLIVEUS = missing.

DCBIRTH (Country Of Birth)

Derived from A-1. If A-1 = “UNITED STATES (50 STATES OR DC),” then DCBIRTH = 1. Otherwise, DCBIRTH = 2.

DMCBIRTH (Mother's Country Of Birth)

Derived from G-1. If G-1 = “UNITED STATES (50 STATES OR DC),” then DMCBIRTH = 1; if G-1 = "99999999999999999999999999999999999999999999997/99999999999999999999999999999999999999999999998", then DMCBIRTH = missing. Otherwise, DMCBIRTH = 2.

DFCBIRTH (Father's Country Of Birth)

Derived from G-3. If G-3 = "UNITED STATES (50 STATES OR DC)", then DFCBIRTH = 1; if G-3 = “99999999999999999999999999999999999999999999997/99999999999999999999999999999999999999999999998,” then DFCBIRTH = missing. Otherwise, DFCBIRTH = 2.

D1STLAN (Language Spoken Before School: 5 Categories)

Derived from A-6. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables when "Other, Specify" was reported in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then D1STLAN = 1; if the respondent speaks both English and Spanish, regardless whether he/she also speaks another language(s), D1STLAN = 2; if the respondent speaks English and another language(s) but not Spanish, D1STLAN = 3; if the respondent speaks Spanish only or Spanish plus another language(s) but not English, D1STLAN = 4; if the respondent speaks neither English nor Spanish, D1STLAN = 5.

D1STLANC (Language Spoken Before School: 3 Categories)

Derived from A-6. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables when "Other, Specify" was reported in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then D1STLANC = 1; if the respondent speaks English and another language(s), D1STLANC = 2; if the respondent does not speak English, D1STLANC = 3.

DHMLANG (Language Spoken At Home When Growing Up: 5 Categories)

Derived from A-5. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables when "Other, Specify" was reported in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then DHMLANG = 1; if the respondent speaks both English and Spanish, regardless whether he/she also speaks another language(s), DHMLANG = 2; if the respondent speaks English and another language(s) but not Spanish, DHMLANG = 3; if the respondent speaks Spanish only or Spanish plus another language(s) but not English, DHMLANG = 4; if the respondent speaks neither English nor Spanish, DHMLANG = 5.

DHMLANGC (Language Spoken At Home When Growing Up: 3 Categories)

Derived from A-5. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables for "Other, Specify" in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then DHMLANGC = 1; if the respondent speaks English and another language(s), DHMLANGC = 2; if the respondent does not speak English, DHMLANGC = 3.

DLANGRW (Language First Learned To Read And Write: 3 Categories)

Derived from A-7. If A-7 = "English," then DLANGRW = 1; if A-7 = "Spanish," then DLANGRW = 2; if A-7 = "99999999999999999999999999999999999997/99999999999999999999999999999999999998," then DLANGRW = missing; if A-7 = missing, then DLANGRW = 1 (these are respondents who reported "English only" to A-5 and A-6). Otherwise, DLANGRW = 3.

DLANGRWC (Language First Learned To Read And Write: 2 Categories)

Derived from A-7. If A-7 = "English," then DLANGRWC = 1; if A-7 = "99999999999999999999999999999999999997/99999999999999999999999999999999999998," then DLANGRWC = missing; if A-7 = missing, then DLANGRWC = 1 (these are respondents who reported "English only" to A-5 and A-6). Otherwise, DLANGRWC = 2.

DEDBFUS (Education Before Coming To The U.S.)

Derived from A-4. If A-4 = 1/2, then DEDBFUS = 1; if A-4 = 3, then DEDBFUS = 2; if A-4 = 4/5/6, then DEDBFUS = 3; if A-4 = missing, then DEDBFUS = 98; if A-4 = 97/98, then DEDBFUS = missing.

DEDATTN (Educational Attainment: 9 Categories)

Derived from B-1. Recoded as follows: If B-1 = 1, then DEDATTN = 1; if B-1 = 2/3, then DEDATTN = 2; if B-1 = 4, then DEDATTN = 3; if B-1 = 5, then DEDATTN = 4; if B-1 = 6, then DEDATTN = 5; if B-1 = 7, then DEDATTN = 6; if B-1 = 8/9, then DEDATTN = 7; if B-1 = 10, then DEDATTN = 8; if B-1 = 11/12, then DEDATTN = 9; if B-1 = 97/98, then DEDATTN = missing.

DEDATTNC (Educational Attainment: 6 Categories)

Derived from B-1. Recoded as follows: If B-1 = 1, then DEDATTNC = 1; if B-1 = 2/3, then DEDATTNC = 2; if B-1 = 4, then DEDATTNC = 3; if B-1 = 5, then DEDATTNC = 4; if B-1 = 6/7/8/9, then DEDATTNC = 5; if B-1 = 10/11/12, then DEDATTNC = 6; if B-1 = 97/98, then DEDATTNC = missing.

DHSAGE (Age Upon Graduation From High School Or Obtaining A Ged)

Derived from A-3, B-1 and B-2. DHSAGE was assigned only to respondents who completed high school/obtained a GED: B-1 > 3. If B-1 < = 3, then DHSAGE = 98.

For respondents with B-1 >3:

- Date of graduation was set as June 30 in the year indicated in B-2.
- Respondent's date of birth (A-3) was subtracted from date of graduation

- This yielded an age expressed in years and months (e.g. 18.5 = 18 years, 6 months). Ages not expressed in whole numbers were rounded down.

- Ages were then recoded into reporting categories 1 and 2. The lower boundary for category 1 was 12 years of age.

If B-1 = 97/98 or if A-3 = 99999997/99999998 or if B-2 = 9997/9998, then DHSAGE = missing.

DMED (Mother's Educational Attainment: 8 Categories)

Derived from G-2 as follows: if G-2 = 1/2, then DMED = 1; if G-2 = 3, then DMED = 2; if G-2 = 4, then DMED = 3; if G-2 = 5, then DMED = 4; if G-2 = 6, then DMED = 5; if G-2 = 7/8, then DMED = 6; if G-2 = 9, then DMED = 7; if G-2 = 10/11, then DMED = 8; if G-2 = 97/98, then DMED = missing.

DMEDC (Mother's Educational Attainment: 5 Categories)

Recoded from G-2 as follows: if G-2 = 1/2, then DMEDC = 1; if G-2 = 3, then DMEDC = 2; if G-2 = 4, then DMEDC = 3; if G-2 = 5/6/7/8, then DMEDC = 4; if G-2 = 9/10/11, then DMEDC = 5; if G-2 = 97/98, then DMEDC = missing.

DFED (Father's Educational Attainment: 8 Categories)

Derived from G-4 as follows: if G-4 = 1/2, then DFED = 1; if G-4 = 3, then DFED = 2; if G-4 = 4, then DFED = 3; if G-4 = 5, then DFED = 4; if G-4 = 6, then DFED = 5; if G-4 = 7/8, then DFED = 6; if G-4 = 9, then DFED = 7; if G-4 = 10/11, then DFED = 8; if G-4 = 97/98, then DFED = missing.

DFEDC (Father's Educational Attainment: 5 Categories)

Recoded from G-4 as follows: if G-4 = 1/2, then DFEDC = 1; if G-4 = 3, then DFEDC = 2; if G-4 = 4, then DFEDC = 3; if G-4 = 5/6/7/8, then DFEDC = 4; if G-4 = 9/10/11, then DFEDC = 5; if G-4 = 97/98, then DFEDC = missing.

DWEEKWG (Weekly Wage: Previous Week)

Derived from D-1, D-3, D-3A and D-4. Note: DWEEKWG was assigned to full-time employees only. If a respondent was not employed full time, DWEEKWG = 98. To determine if a respondent was employed full time, use responses to D-1A, D-1C, D-1E and D-1F. If a respondent indicated in D-1A, D-1C, D-1E and D-1F that she/he was employed, she/he was employed full time. For full time employees:

- If reported pay was gross pay, i.e., if D-3A = 2, then reported income was converted to weekly gross pay as follows:

  if D-3 (Unit) = 1, then weekly gross pay = D-3 (Dollar amount)*40;

  if D-3 (Unit) = 2, then weekly gross pay = D-3 (Dollar amount)*5;

if D-3 (Unit) = 3, then weekly gross pay = D-3 (Dollar amount);

if D-3 (Unit) = 4, then weekly gross pay = D-3 (Dollar amount)/2;

if D-3 (Unit) = 5, then weekly gross pay = D-3 (Dollar amount)/4.3;

if D-3 (Unit) = 6, then weekly gross pay = D-3 (Dollar amount)/52;

if D-3 (Unit) = 91, then do as follows:

If "Other, specify" of D-3 (Unit) = 'EVERY 15TH AND 31'/'1700 FOR 15 DAY PERIOD (GET PAID 2 A MTH'/'TWICE PER MO'/'TWICE A MONTH'/'1ST AND 15TH'/'2XMO'/'2XMO.'/ 'BIMONTHLY'/'BI MONTHLY', then weekly gross pay = D-3 (Dollar amount)/(4.3*0.5).

- If reported pay was take home rather than gross, i.e., if D-3A = 1, reported income was converted to weekly gross pay in two steps.

Step 1: Convert reported income to weekly take home pay as follows:

if D-3 (Unit) = 1, then weekly take home pay = D-3 (Dollar amount)*40;

if D-3 (Unit) = 2, then weekly take home pay = D-3 (Dollar amount)*5;

if D-3 (Unit) = 3, then weekly take home pay = D-3 (Dollar amount);

if D-3 (Unit) = 4, then weekly take home pay = D-3 (Dollar amount)/2;

if D-3 (Unit) = 5, then weekly take home pay = D-3 (Dollar amount)/4.3;

if D-3 (Unit) = 6, then weekly take home pay = D-3 (Dollar amount)/52;

if D-3 (Unit) = 91, then do as follows:

If "Other, specify" of D-3 (Unit) = 'EVERY 15TH AND 31'/'1700 FOR 15 DAY PERIOD (GET PAID 2 A MTH'/'TWICE PER MO'/'TWICE A MONTH'/'1ST AND 15TH'/'2XMO'/'2XMO.'/'BIMONTHLY'/'BI MONTHLY', then weekly take home pay = D-3 (Dollar amount)/(4.3*0.5).

Step 2: Add the following tax withholding adjustments to weekly take home pay to estimate weekly gross pay:

1. Add FICA adjustment (Social and Medicare) at a flat rate of 7.65 percent.
2. Add adjustment based on IRS withholding tables for single taxpayers in 2003 (IRS Form Pub15-T, Table 1 (Weekly Payroll Period), (a) Single Person).
3. Add proxy adjustment for state taxes and miscellaneous deductions at a rate of 10 percent.

- If D-3A = missing or 7, then weekly gross pay = missing.

Continuous weekly gross pay was then rounded and recoded to categorical DWEEKWG.

DINCOME (Income Adequacy)

Derived for household sample only from CALCAGEA - CALCAGEY from the Screener, and H-1, H-2 and K-2 from the Background Questionnaire, using the table "Poverty Thresholds for 2003 by Size of Family and Number of Related Children Under 18 Years (Dollars)" published by the Census Bureau. Steps:

1. Use CALCAGEA - CALCAGEY to determine the total number of people in the home, the number of related children under age 18 in the home, and the number of people age 65 and over in the home. The combined information would determine the "Family type" in the Poverty Thresholds Table. Note: When there were two persons in the household, the following guidelines were used to determine who the householder was and hence the "family type":

    i. If neither of the 2 persons was a child, and both people were under 65, then family type = "Two persons, Householder under 65, No children."

    ii. If 1 person was a child and the other person was under 65, then family type = "Two persons, Householder under 65, 1 child."

    iii. If neither of the 2 persons was a child, and 1 person was over 65 and the other person was under 65, then family type = "Two persons, Householder 65 and over, No children."

    iv. If 1 of the 2 persons was a child, and the other person was over 65, then family type = "Two persons, Householder 65 and over, 1 child."

2. Compare the lower boundary of the income range reported in K-2 with the poverty thresholds in the Poverty Thresholds Table for the appropriate family type. If the lower boundary of K-2 was less than the corresponding poverty threshold, then DINCOME = 1. Otherwise, DINCOME = 2.

Note: If K-2 = 97/98, use the follow-up probes to identify the range of income by first recoding them into the following categories:

1 = <$10k
2 = $10k – $15k
3 = <$15k
4 = $15k – $30k
5 = <$20k
6 = $20k – $30k
7 = <$30k
8 = $30k – $40k
9 = $30k – $60k
10 = <$40k
11 = $40k – $60k
12 = <$60k
13 = $60k – $100k
14 = <$100k
15 = over $30k
16 = over $60k
17 = over $100k

Categories 5 (< 20k), 7 (< 30k), 10 (< 40k), 12 (< 60k), and 14 (< 100k) were treated as missing. Then the lower boundaries of the above income categories were compared to the poverty thresholds corresponding to the appropriate family type to create DINCOME.

DMARITAL (Marital Status)

Derived from I-1. If I-1 = 1, then DMARITAL = 1; if I-1 = 2/3/4, then DMARITAL = 2; if I-1 = 5/6, then DMARITAL = 3; if I-1 = 7/8, then DMARITAL = missing.

DLFORCE (Labor Force Participation)

Derived from D-1 and D-2. If D-1B = 2, then DLFORCE = 2. Otherwise, if D-1A = 1 or D-1C = 3, then DLFORCE = 1. Otherwise, if D-1E = 5 or D-1F = 6, then DLFORCE = 3. Otherwise, if D-1D = 4 and D-2 = 1, then DLFORCE = 4. Otherwise, if D-1D = 4 and D-2 = 2, then DLFORCE = 5. Otherwise, if D-1G = 7 or D-1H = 8 or D-1I = 9 or D-1J = 10 or D-1K = 91, then DLFORCE = 5; all else DLFORCE = missing.

DWFTIME (Length Of Participation In Welfare Programs)

Derived from I-3G, I-8D, I-6, and I-10. If I-8D = 2, then DWFTIME = 1. Otherwise, if I-3G = 2/7/8 and I-8D = 7/8, then DWFTIME = missing. Otherwise, if I-6 = 1/2/3 or I-10 = 1/2/3, then DWFTIME = 2; if I-6 = 4/5 or I-10 = 4/5, then DWFTIME = 3; if I-6 = 7/8 or I-10 = 7/8, then DWFTIME = missing.

DVOTE (Voting In The Most Recent Presidential Election)

Derived from C-8, C-10 and C-11. If C-8 = 2, DVOTE = 0. (Note: Only respondents who were not born in the U.S. were asked this question. Everyone born in the U.S. was assumed to be a citizen.) Otherwise, If C-10 = 3 (voted), DVOTE = 2. Otherwise, if C-10 = 4, then DVOTE = 1; if C-10 = 3, then DVOTE = 2; also, If C-10 = 1 and C-11 = 1, then DVOTE = 2; if C-10 = 2, then DVOTE = 3; if C-10 = 1 and C-11 = 2, then DVOTE = 1. Otherwise, DVOTE = missing.

DEMPTYPC (Type Of Employer In The Past Three Years: 3 Categories)

Recoded from D-13. If D-13 = 2, then DEMPTYPC = 1; if D-13 = 3, then DEMPTYPC = 2; if D-13 = 1/4, then DEMPTYPC = 3; if D-13 = missing, then DEMPTYPC = 98; if D-13 = 7/8, then DEMPTYPC = missing.

DSPUDSTD (How Well Understand Spanish)

Derived from A-14A and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. A-14A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPUDSTD, all non-English languages were checked and if the non-English language = "SPANISH," then DSPUDSTD = the A-14A that was linked to that SPANISH language. If A-14A = 7/8, then DSPUDSTD was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12 and A-13, DSPUDSTD = 98.

DOTUDSTD (How Well Understand Other Non-English Language)

Derived from A-14A and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. A-14A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTUDSTD, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12 and A-13, DOTUDSTD = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTUDSTD = the A-14A that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTUDSTD = the A-14A that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTUDSTD = "Well." If the linked A-14A = 7/8, then DOTUDSTD was recoded as missing.

DSPSPEAK (How Well Speak Spanish)

Derived from A-14B and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. A-14B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPSPEAK, all non-English languages were checked and if the non-English language = "SPANISH," then DSPSPEAK = the A-14B that was linked to that SPANISH language. If A-14B = 7/8, then DSPSPEAK was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPSPEAK = 98.

DOTSPEAK (How Well Speak Other Non-English Language)

Derived from A-14B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTSPEAK, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTSPEAK = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTSPEAK = the A-14B that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTSPEAK = the A-14B that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTSPEAK = "Well." If the linked A-14B = 7/8, then DOTSPEAK was recoded as missing.

DSPREAD (How Well Read Spanish)

Derived from A-14C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPREAD, all non-English languages were checked and if the non-English language = "SPANISH," then DSPREAD = the A-14C that was linked to that SPANISH language. If A-14C = 7/8, then DSPREAD was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPREAD = 98.

DOTREAD (How Well Read Other Non-English Language)

Derived from A-14C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTREAD, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTREAD = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTREAD = the A-14C that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTREAD = the A-14C that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTREAD = "Well." If the linked A-14C = 7/8, then DOTREAD was recoded as missing.

DSPWRITE (How Well Write Spanish)

Derived from A-14D and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14D was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPWRITE, all non-English languages were checked and if the non-English language = "SPANISH," then DSPWRITE = the A-14D that was linked to that SPANISH language. If A-14D = 7/8, then DSPWRITE was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPWRITE = 98.

DOTWRITE (How Well Write Other Non-English Language)

Derived from A-14D and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14D was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTWRITE, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTWRITE = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTWRITE = the A-14D that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTWRITE = the A-14D that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTWRITE = "Well." If the linked A-14D = 7/8, then DOTWRITE was recoded as missing.

DSPINFO (How Much Info Got In Spanish About Current Events/Public Affairs/Government)

Derived from C-2 and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. C-2 was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPINFO, all non-English languages were checked and if the non-English language = "SPANISH," then DSPINFO = the C-2 that was linked to that SPANISH language. If C-2 = 7/8, then DSPINFO was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPINFO = 98.

DOTINFO (How Much Info Got In Other Non-English Language About Current Events/Public Affairs/Government)

Derived from C-2 and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. C-2 was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTINFO, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, DOTINFO = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTINFO = the C-2 that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTINFO = the C-2 that was linked to the language in which the respondent got the most information. For example, if a respondent spoke two non-English non-Spanish languages and he/she got "most" info in one language and "some" info in the other, then DOTINFO = "Most." If the linked C-2 = 7/8, then DOTINFO was recoded as missing.

DCLANGS (Language Usually Speak Now: 3 Categories)

Derived from A-11. If A-11 = missing, then DCLANGS = 98; if A-11 = "99999999999999999999999999999999999997 or 99999999999999999999999999999999999998," then DCLANGS = missing; if A-11 = "English," then DCLANGS = 1; if A-11 = "Spanish," then DCLANGS = 2; otherwise DCLANGS = 3.

DCLANGSC (Language Usually Speak Now: 2 Categories)

Derived from A-11. If A-11 = "English," then DCLANGSC = 1; if A-11 = "99999999999999999999999999999999999997 or 99999999999999999999999999999999999998," then DCLANGSC = missing; otherwise, DCLANGSC = 2.

DOLANGSB (Other Language Speak Best)

Derived from A-13. If A-13 = "Spanish," then DOLANGSB = 1; if A-13 = missing, then DOLANGSB = 98; otherwise, DOLANGSB = 2.

DDTYPE (Type Of High School Degree: 4 Categories)

Derived from B-3S. If B-3S = missing, then DDTYPE = 98; if B-3S = 1/2, then DDTYPE = 1; if B-3S = 3, then DDTYPE = 2; if B-3S = 4, then DDTYPE = 3; otherwise, if B-3S = 6/97/98, then DDTYPE = 4.

DDTYPEC (Type Of High School Degree: 3 Categories)

Derived from B-3S. If B-3S = missing, then DDTYPEC = 98; if B-3S = 1/2, then DDTYPEC = 1; if B-3S = 4, then DDTYPEC = 2; otherwise, if B-3S = 3/6/99, then DDTYPEC = 3.

DSGRDHS (What State Did You Live In When You Graduated High School)

Derived from B-5. If B-5_state = missing, then DSGRDHS = 98; if B-5_state = "99999999999999999997 or 99999999999999999998," then DSGRDHS = missing; if B-5_state = "state," then DSGRDHS = 1; if B-5_state not equal to "state," then DSGRDHS = 2; otherwise, if B-5_state = "outside US," then DSGRDHS = 2.

DSGRDCO (What State Did You Live In When You Graduated College)

Derived from B-8. If B-8_state = missing, then DSGRDCO = 98; if B-8_state = "99999999999999999997 or 99999999999999999998," then DSGRDCO = missing; if B-8_state = "state," then DSGRDCO = 1; if B-8_state not equal to "state," then DSGRDCO = 2; if B-8_STATE = "outside US," then DSGRDCO = 2.

DOLSOPT (Other Language Often Spoken Combined: 5 Categories)

Derived from A-11 and A-12. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables when "Other, Specify" was reported in the 5 initial variables. All 10 variables were cross-checked and recoded in combination with A-11 as follows: If the respondent speaks English only, then DOLSOPT = 1; If the respondent speaks both English and Spanish, regardless whether he/she also speaks another language(s), DOLSOPT = 2; If the respondent speaks English and another language(s) but not Spanish, DOLSOPT = 3; If the respondent speaks Spanish only or Spanish plus another language(s) but not English, DOLSOPT = 4; If the respondent speaks neither English nor Spanish, DOLSOPT = 5.

DSPPAPER (How Often Read Newspapers/Magazines In Spanish)

Derived from E-3A and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. E-3A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPPAPER, all non-English languages were checked and if the non-English language = "SPANISH," then DSPPAPER = the E-3A that was linked to that SPANISH language. If E-3A = 7/8, then DSPPAPER was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPPAPER = 98.

DOTPAPER (How Often Read Newspapers/Magazines In Other Non-English Language)

Derived from E-3A and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-3A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTPAPER, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTPAPER = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTPAPER = the E-3A that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTPAPER = the E-3A that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTPAPER = "Everyday," If the linked E-3B = 7/8, then DOTPAPER was recoded as missing.

DSPBOOK (How Often Read Books In Spanish)

Derived from E-3B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-3B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPBOOK, all non-English languages were checked and if the non-English language = "SPANISH," then DSPBOOK = the E-3B that was linked to that SPANISH language. If E-3B = 7/8, then DSPBOOK was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPBOOK = 98.

DOTBOOK (How Often Read Books In Other Non-English Language)

Derived from E-3B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-3B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTBOOK, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTBOOK = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTBOOK = the E-3B that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTBOOK = the E-3B that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTBOOK = "Everyday." If the linked E-3B = 7/8, then DOTBOOK was recoded as missing.

DSPNOTES (How Often Read Notes In Spanish)

Derived from E-3C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-3C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPNOTES, all non-English languages were checked and if the non-English language = "SPANISH," then DSPNOTES = the E-3C that was linked to that SPANISH language. If E-3C = 7/8, then DSPNOTES was recoded as missing. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPNOTES = 98.

DOTNOTES (How Often Read Notes In Other Non-English Language)

Derived from E-3C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-3C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTNOTES, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTNOTES = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTNOTES = the E-3C that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTNOTES = the E-3C that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTNOTES = "Everyday." If the linked E-3C = 7/8, then DOTNOTES was recoded as missing.

DRFSSCHC (Reason For Stopping School Before College Degree)

Derived from B-4. If B-4 = missing, then DRFSSCHC = 98; if B-4 = 1, then DRFSSCHC = 98; if B-4 = 97/98, then DRFSSCHC = missing; if B-4 = 2, then DRFSSCHC = 1; if B-4 = 3, then DRFSSCHC = 2; if B-4 = 4, then DRFSSCHC = 3; if B-4 = 5, then DRFSSCHC = 4; if B-4 = 6/7, then DRFSSCHC = 5; if B-4 = 8/9/10/11, then DRFSSCHC = 6; if B-4 = 13, then DRFSSCHC = 7; if B-4 = 12, then DRFSSCHC = 8; otherwise, if B-4 = 14, then DRFSSCHC = 8.

DBQ1435 (Frequency Of Volunteering)

Derived from C-3 and C-4. If C-3 = 2, then DBQ1435 = 5. Otherwise, if C-4 = missing, then DBQ1435 = 98; otherwise, if C-4 = 7/8, then DBQ1435 = missing.

DBQ1530B (Not Employed Because Retired)

Derived from D-6A and D-6B. If D-6A = 7/8, then DBQ1530B = missing; if D-6B = missing, then DBQ1530B = 98; otherwise, DBQ1530B = D-6B.

DBQ1530C (Not Employed Because Taking Care Of Home Or Family)

Derived from D-6A and D-6C. If D-6A = 7/8, then DBQ1530C = missing; if D-6C = missing, then DBQ1530C = 98; otherwise, DBQ1530C = D-6C.

DBQ1530D (Not Employed Because Going To School)

Derived from D-6A and D-6D. If D-6A = 7/8, then DBQ1530D = missing; if D-6D = missing; then DBQ1530D = 98; otherwise, DBQ1530D = D-6D.

DBQ1530E (Not Employed Because Could Not Find Work)

Derived from D-6A and D-6E. If D-6A = 7/8, then DBQ1530E = missing; if D-6E = missing, then DBQ1530E = 98; otherwise, DBQ1530E = D-6E.

DBQ1530F (Not Employed Because Of Other Reason)

Derived from D-6A and D-6F. If D-6A = 7/8, then DBQ1530F = missing; if D-6F = missing, then DBQ1530F = 98; otherwise, DBQ1530F = D-6F.

DUSECOMP (Ever Use A Computer)

Derived from C-1C and E-1. If C-1C = 1/2/3, then DUSECOMP = 1. Otherwise, if E-1 = 7/8/missing, then DUSECOMP = missing. Otherwise, DUSECOMP = E-1.

DREDCHLD (How Many Times Read To Child: Past Week)

Derived from H-5 and H-6. If H-5 = 2, then DREDCHLD = 5. Otherwise, if H-6 = 7, then DREDCHLD = missing; if H-6 = missing, then DREDCHLD = 98. Otherwise, DREDCHLD = H-6.

DBQ2165 (Ever Received Supplemental Security Income)

Derived from I-8A and I-3B. If I-8A = 1 or I-3B = 1/3, then DBQ2165 = 1; if I-8A = 2, then DBQ2165 = 2. Otherwise, if I-8A = 7/8/missing, then DBQ2165 = missing.

DBQ2170 (Ever Received Food Stamps)

Derived from I-8B and I-3D. If I-8B = 1 or I-3D = 1/3, then DBQ2170 = 1; if I-8B = 2, then DBQ2170 = 2. Otherwise, if I-8B = 7/8/missing, then DBQ2170 = missing.

DBQ2175 (Ever Received Wic Supplemental Nutrition Benefits)

Derived from I-8C and I-3E. If I-8C = 1 or I-3E = 1/3, then DBQ2175 = 1; if I-8C = 2, then DBQ2175 = 2. Otherwise, if I-8C = 7/8/missing, then DBQ2175 = missing.

DBQ2180 (Ever Received Tanf Public Assistance Or Public Welfare Payments)

Derived from I-8D and I-3G. If I-8D = 1 or I-3G = 1/3, then DBQ2180 = 1; if I-8D = 2, then DBQ2180 = 2; if I-8D = 7/8, then DBQ2180 = missing.

DWLFLST (Last Received Welfare Payments: 4 Categories)

Derived from I-3G and I-9. If I-3G = 1/3, then DWLFLST = 1; if I-9 = 1, then DWLFLST = 2; if I-9 = 2, then DWLFLST = 3; if I-9 = 3, then DWLFLST = 4; if I-9 = missing, then DWLFLST = 98; if I-9 = 7/8, then DWLFLST = missing.

DWLFLSTC (Last Received Welfare Payments: 2 Categories)

Derived from I-3G and I-9. If I-3G = 1/3, then DWLFLSTC = 1; also, if I-9 = 1/2, then DWLFLSTC = 1; if I-9 = 3, then DWLFLSTC = 2; if I-9 = missing, then DWLFLSTC = 98; if I-9 = 7/8, then DWLFLSTC = missing.

DBQ2421 (Approximate Personal Income: 8 Categories)

Derived from K-1. if K-1 = 1/14, then DBQ2421 = 1; if K-1 = 2/3, then DBQ2421 = 2; if K-1 = 4/5, then DBQ2421 = 3; if K-2 = 6, then DBQ2421 = 4; if K-2 = 7 then DBQ2421 = 5; if K-2 = 8, then DBQ2421 = 6; if K-1 = 9/10, then DBQ2421 = 7; if K-1 = 11/12/13, then DBQ2421 = 8. Otherwise, DBQ2421 = missing.

DBQ2430 (Approximate Household Income: 8 Categories)

Derived K-2. If K-2 = 1/2/3/15, then DBQ2430 = 1; if K-2 = 4/5, then DBQ2430 = 2; if K-2 = 6, then DBQ2430 = 3; if K-2 = 7, then DBQ2430 = 4; if K-2 = 8, then DBQ2430 = 5; if K-2 = 9/10, then DBQ2430 = 6; if K-2 = 11/12, then DBQ2430 = 7; if K-2 = 13/14, then DBQ2430 = 8. Otherwise, DBQ2430 = missing.

ICODE_C (Industry)

Derived from D-9 and D-10. If D-9 = 3, then ICODE_C = 21. Otherwise, use D-10. Responses to D-10 were categorized into the standard 4-digit classifications used by the U.S. Census. These 4-digit classifications were further categorized into the Census standard combinations of 2-digit classifications with one exception: “Military” was combined into “Unknown (Missing).”

OCODE_C (Occupation)

Derived from D-9 and D-11. If D-9 = 3, then OCODE_C = 31. Otherwise, use D-11. Responses to D-11 were categorized into the standard 4-digit classifications used by the U.S. Census. These 4-digit classifications were further categorized into the Census standard combinations of 2-digit classifications with two exceptions: “Military” was combined into “Unknown (Missing)” and “Funeral workers” was combined into “Personal Care and Service.”

2. Derived variables for the NAAL 2003 prison sample

DAGEC (Age: 4 Categories)

Recoded from a continuous age variable derived from date of interview and the BQ response for date of birth, A-3.

DRACE (Race/Ethnicity)

Derived from J-1 and J-3. If J-1 = 1, then DRACE = 3. Otherwise, if J-3A = 1 and J-3B-E = missing, then DRACE = 1; if J-3B = 1 and (J-3A and J-3C-E = missing), then DRACE = 2. Otherwise, DRACE = 4.

Note: White, Black and Other categories include no Hispanics.

DARRIVE (Age Arrived In The U.S.)

Derived from A-1 and A-1A. If A-1 = "UNITED STATES (50 STATES OR DC)," then DARRIVE = 1. Otherwise, recode continuous responses for A-1A to categorical. If A-1A = missing, then DARRIVE = missing.

DENGAGE (Age Learned To Speak English)

Derived from A-8. If A-8 = missing, then DENGAGE = 1. These are people who speak only English. Otherwise, if A-8 = 1/2, then DENGAGE = 2; if A-8 = 3/4/5, then DENGAGE = 3; if A-8 = 95, then DENGAGE = 4; if A-8 = 99, then DENGAGE = missing.

DCBIRTH (Country Of Birth)

Derived from A-1. If A-1 = "UNITED STATES (50 STATES OR DC)," then DCBIRTH = 1. Otherwise, DCBIRTH = 2.

DMCBIRTH (Mother's Country Of Birth)

Derived from G-1. If G-1 = "UNITED STATES (50 STATES OR DC)," then DMCBIRTH = 1; if G-1 = "99999999999999999999999999999999999999999999999999," then DMCBIRTH = missing. Otherwise, DMCBIRTH = 2.

DFCBIRTH (Father's Country Of Birth)

Derived from G-3. If G-3 = "UNITED STATES (50 STATES OR DC)," then DFCBIRTH = 1; if G-3 = "99999999999999999999999999999999999999999999999999," then DFCBIRTH = missing. Otherwise, DFCBIRTH = 2.

D1STLANC (Language Spoken Before School: 3 Categories)

Derived from A-6. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables when "Other, Specify" was reported in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then D1STLANC = 1; if the respondent speaks English and another language(s), D1STLANC = 2; if the respondent does not speak English, D1STLANC = 3.

DHMLANGC (Language Spoken At Home When Growing Up: 3 Categories)

Derived from A-5. Note: Languages that all apply were selected from lookup table and reported in 5 initial variables and 5 follow-up variables for "Other, Specify" in the 5 initial variables. All 10 variables were cross-checked and recoded as follows: If the respondent speaks English only, then DHMLANGC = 1; if the respondent speaks English and another language(s), DHMLANGC = 2; if the respondent does not speak English, DHMLANGC = 3.

DLANGRWC (Language First Learned To Read And Write: 2 Categories)

Derived from A-7. If A-7 = "English," then DLANGRWC = 1; if A-7 = missing, then DLANGRWC = 1 (these are respondents who reported "English only" to A-5 and A-6). Otherwise, DLANGRWC = 2.

DEDBFUS (Education Before Coming To The U.S.)

Derived from A-4. If A-4 = 1/2, then DEDBFUS = 1; if A-4 = 3, then DEDBFUS = 2; if A-4 = 4/5/6, then DEDBFUS = 3; if A-4 = missing, then DEDBFUS = 98.

DEDATTNC (Educational Attainment: 6 Categories)

Derived from B-1, B-2 and B-5. Note: Education was recorded before (B-1) and while (B-5) in prison. If B-2 = 1 and B-5 >B-1, then use B-5 for DEDATTNC. If B-2 ~ = 1, then use B-1 for DEDATTNC. B-5/B-1 was recoded into corresponding categories of DEDATTNC.

DHSAGE (Age Upon Graduation From High School Or Obtaining A Ged)

Derived from A-3, B-1 and B-6. DHSAGE was assigned only to respondents who completed high school/obtained a GED: B-1 >3. If B-1 < = 3, then DHSAGE = 98.

For respondents with B-1 >3:

- Date of graduation was set as June 30 in the year indicated in B-6.
- Respondent's date of birth (A-3) was subtracted from date of graduation.

- This yielded an age expressed in years and months (e.g. 18.5 = 18 years, 6 months). Ages not expressed in whole numbers were rounded down.

- Ages were then recoded into reporting categories 1 and 2. The lower boundary for category 1 was 12 years of age.

DMEDC (Mother's Educational Attainment: 5 Categories)

Recoded from G-2 as follows: if G-2 = 1/2, then DMEDC = 1; if G-2 = 3, then DMEDC = 2; if G-2 = 4, then DMEDC = 3; if G-2 = 5/6/7/8, then DMEDC = 4; if G-2 = 9/10/11, then DMEDC = 5; if G-2 = 99, then DMEDC = missing.

DFEDC (Father's Educational Attainment: 5 Categories)

Recoded from G-4 as follows: if G-4 = 1/2, then DFEDC = 1; if G-4 = 3, then DFEDC = 2; if G-4 = 4, then DFEDC = 3; if G-4 = 5/6/7/8, then DFEDC = 4; if G-4 = 9/10/11, then DFEDC = 5; if G-4 = 99, then DFEDC = missing.

DMARITAL (Marital Status)

Derived from H-1. If H-1 = 1, then DMARITAL = 1; if H-1 = 2/3, then DMARITAL = 2; if H-1 = 4/5, then DMARITAL = 3; if H-1 = 7/8, then DMARITAL = missing.

DWFTIME (Length Of Participation In Welfare Programs)

Derived from H-2D and H-4. If H-2D = 2, then DWFTIME = 1. Otherwise, if H-4 = 1/2/3, then DWFTIME = 2; if H-4 = 4/5, then DWFTIME = 3. Otherwise, DWFTIME = missing.

DVOTE (Voting In The Most Recent Presidential Election)

Derived from E-9, E-11 and E-12. If E-9 = 2, DVOTE = 0. (Note: Only respondents who were not born in the U.S. were asked this question. Everyone born in the U.S. was assumed to be a citizen.) Otherwise, if E-11 = 3 (voted), DVOTE = 2.

Otherwise, if both E-11 and E-12 = missing, then DVOTE = 98. These were prisoners who were in prison for current offense in November 2000 and skipped E-11 and E-12.

Otherwise, if E-11 = 4, then DVOTE = 1; if E-11 = 3, then DVOTE = 2; also, If E-11 = 1 and E-12 = 1, then DVOTE = 2; if E-11 = 2, then DVOTE = 3; if E-11 = 1 and E-12 = 2, then DVOTE = 1. Otherwise, DVOTE = missing.

DEMPTYPC (Type Of Employer In The Past Three Years: 3 Categories)

Derived from D-12. If D-12 = 2, then DEMPTYPC = 1; if D-12 = 3, then DEMPTYPC = 2; if D-12 = 1/4, then DEMPTYPC = 3; if D-12 = missing, then DEMPTYPC = 98; if D-12 = 8, then DEMPTYPC = missing.

DSPUDSTD (How Well Understand Spanish)

Derived from A-14A and the associated listings of the non-English languages identified in A-6, A-11, A-12 and A-13. Note that respondents were allowed to select multiple non-English languages. A-14A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPUDSTD, all non-English languages were checked and if the non-English language = "SPANISH," then DSPUDSTD = the A-14A that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPUDSTD = 98.

DOTUDSTD (How Well Understand Other Non-English Language)

Derived from A-14A and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTUDSTD, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTUDSTD = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTUDSTD = the A-14A that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTUDSTD = the A-14A that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTUDSTD = "Well."

DSPSPEAK (How Well Speak Spanish)

Derived from A-14B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPSPEAK, all non-English languages were checked and if the non-English language = "SPANISH," then DSPSPEAK = the A-14B that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPSPEAK = 98.

DOTSPEAK (How Well Speak Other Non-English Language)

Derived from A-14B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTSPEAK, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTSPEAK = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTSPEAK = the A-14B that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTSPEAK = the A-14B that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTSPEAK = "Well."

DSPREAD (How Well Read Spanish)

Derived from A-14C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPREAD, all non-English languages were checked and if the non-English language = "SPANISH," then DSPREAD = the A-14C that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPREAD = 98.

DOTREAD (How Well Read Other Non-English Language)

Derived from A-14C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTREAD, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTREAD = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTREAD = the A-14C that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTREAD = the A-14C that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTREAD = "Well."

DSPWRITE (How Well Write Spanish)

Derived from A-14D and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14D was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPWRITE, all non-English languages were checked and if the non-English language = "SPANISH," then DSPWRITE = the A-14D that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, A-11, A-12, and A-13, DSPWRITE = 98.

DOTWRITE (How Well Write Other Non-English Language)

Derived from A-14D and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. A-14D was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTWRITE, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, A-11, A-12, and A-13, DOTWRITE = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTWRITE = the A-14D that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTWRITE = the A-14D that was linked to the language rated as best. For example, if a respondent spoke two non-English non-Spanish languages and if one language was rated as "Well" and the other as "Not well," then DOTWRITE = "Well."

DSPINFO (How Much Info Got In Spanish About Current Events/Public Affairs/Government)

Derived from E-2 and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-2 was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPINFO, all non-English languages were checked and if the non-English language = "SPANISH," then DSPINFO = the E-2 that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPINFO = 98.

DOTINFO (How Much Info Got In Other Non-English Language About Current Events/Public Affairs/Government)

Derived from E-2 and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. E-2 was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTINFO, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, DOTINFO = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTINFO = the E-2 that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTINFO = the E-2 that was linked to the language in which the respondent got the most information. For example, if a respondent spoke two non-English non-Spanish languages and he/she got "most" info in one language and "some" info in the other, then DOTINFO = "Most,"

DCLANGSC (Language Usually Speak Now: 2 Categories)

Derived from A-11. If A-11 = "English," then DCLANGSC = 1; otherwise, DCLANGSC = 2.

DOLANGSB (Other Language Speak Best)

Derived from A-13. If A-13 = "Spanish," then DOLANGSB = 1; if A-13 = missing, then DOLANGSB = 98; otherwise, DOLANGSB = 2.

DDTYPEC (Type Of High School Degree: 3 Categories)

Derived from B-3S. If B-3S = missing, then DDTYPEC = 98; if B-3S = 1/2, then DDTYPEC = 1; if B-3S = 4, then DDTYPEC = 2; otherwise, if B-3S = 3/6/99, then DDTYPEC = 3.

DSPPAPER (How Often Read Newspapers/Magazines In Spanish)

Derived from F-2A and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPPAPER, all non-English languages were checked and if the non-English language = "SPANISH," then DSPPAPER = the F-2A that was linked to the SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPPAPER = 98.

DOTPAPER (How Often Read Newspapers/Magazines In Other Non-English Language)

Derived from F-2A and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2A was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTPAPER, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTPAPER = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTPAPER = the F-2A that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTPAPER = the F-2A that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTPAPER = "Everyday."

DSPBOOK (How Often Read Books In Spanish)

Derived from F-2B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPBOOK, all non-English languages were checked and if the non-English language = "SPANISH," then DSPBOOK = the F-2B that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPBOOK = 98.

DOTBOOK (How Often Read Books In Other Non-English Language)

Derived from F-2B and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2B was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTBOOK, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTBOOK = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTBOOK = the F-2B that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTBOOK = the F-2B that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTBOOK = "Everyday."

DSPNOTES (How Often Read Notes In Spanish)

Derived from F-2C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DSPNOTES, all non-English languages were checked and if the non-English language = "SPANISH," then DSPNOTES = the F-2C that was linked to that SPANISH language. For respondents who did not speak SPANISH as their non-English language or who were identified as "English only" in A-6, DSPNOTES = 98.

DOTNOTES (How Often Read Notes In Other Non-English Language)

Derived from F-2C and the associated listings of the non-English languages identified in A-6, A-11, A-12, and A-13. Note that respondents were allowed to select multiple non-English languages. F-2C was repeated for each of the non-English languages and therefore comprised an array of responses linked to each of the non-English languages. To create DOTNOTES, all non-English languages were checked. If a respondent spoke SPANISH only as non-English language or was identified as "English only" in A-6, then DOTNOTES = 98. Otherwise, if a respondent spoke only one non-English non-Spanish language, then DOTNOTES = the F-2C that was linked to that language; if a respondent spoke more than one non-English non-Spanish languages, then DOTNOTES = the F-2C that was linked to the language in which the respondent read the most often. For example, if a respondent spoke two non-English non-Spanish languages and if one language was reported as "Everyday" and the other as "Once a week," then DOTNOTES = "Everyday,"

DRFSSCHC (Reason For Stopping School Before College Degree)

Derived from B-4. If B-4 = missing, then DRFSSCHC = 98; if B-4 = 1, then DRFSSCHC = 98; if B-4 = 2, then DRFSSCHC = 1; if B-4 = 3, then DRFSSCHC = 2; if B-4 = 4, then DRFSSCHC = 3; if B-4 = 5, then DRFSSCHC = 4; if B-4 = 6/7, then DRFSSCHC = 5; if B-4 = 8/9/10/11, then DRFSSCHC = 6; if B-4 = 13, then DRFSSCHC = 7; if B-4 = 12, then DRFSSCHC = 8; otherwise, if B-4 = 14, then DRFSSCHC = 8.

DWLFLSTC (Last Received Welfare Payments: 2 Categories)

Derived from H-3. If H-3 = 1/2/3, then DWLFLSTC = 1; if H-3 = 4, then DWLFLSTC = 2; if H-3 = missing, then DWLFLSTC = 98.

DPGED (Ged Earned While In Prison)

Derived from B-1, B-3, B-5, and B-12. If B-1 > = 4 and (B-5 > = 4 or B-5 = missing or B-5 = 9), then DPGED = 1; if B-1 < 4 and B-5 > = 4, or B-1 > = 4 and B-5 = 3, then DPGED = 2. Otherwise, if B-1 <4 and (B-3 = 1 or B-12 = 4), then DPGED = 3. Otherwise, DPGED = 4. Note: There were 5 cases where B-1 > = 4 and (B-5 = 1/2). All 5 cases indicated that they were not currently enrolled in academic (B-3) or basic skills (B-12) classes. For all 5 cases, DPGED = 4.

DPBQ1209 (Currently Enrolled In Academic Classes)

Derived from B-3. If B-3 = 1, then DPBQ1209 = 1; if B-3 = 2, then DPBQ1209 = 2; if B-3 = missing, then DPBQ1209 = 2. Note: Respondents who answered "No" to B-2 skipped B-3 and were coded as missing in B-3.

DPBLOCAT (Where Took Basic Skills Class)

Derived from B-10 and B-11. If B-10 = 2, then DPBLOCAT = 4. If B-10 = 1, then do as follows: if B-11A = 1, then DPBLOCAT = 1; if B-11B = 2, then DPBLOCAT = 2; if B-11C = 3, then DPBLOCAT = 3. Note: Respondents were allowed to select more than 1 answer for B-11. In this situation, because education completed in prison was of most interest, the answer that was closest to the respondent's most recent incarceration was taken for DPBLOCAT. For example: if B-11A = 1 and B-11C = 3, then DPBLOCAT = 1; if B-11A = 1 and B-11B = 2, then DPBLOCAT = 1; if B-11B = 2 and B-11C = 3, then DPBLOCAT = 2.

DPITCERT (Completion Of It Certification While In Prison)

Derived from B-16. If B-16 = 1, then DPITCERT = 1; if B-16 = 2, then DPITCERT = 2; if B-16 = missing, then DPITCERT = 3.

DPOTCERT (Completion Of Other Job Certification While In Prison)

Derived from B-19. If B-19 = 1, then DPOTCERT = 1; if B-19 = 2, then DPOTCERT = 2; if B-19 = missing, then DPOTCERT = 3.

DPVOC (Length Of Time In Prison Vocational Training Program)

Derived from C-2 and C-4. If C-2 = 2, then DPVOC = 0. Otherwise, if C-4 = 1 or 2, then DPVOC = 1; if C-4 = 3, then DPVOC = 2; if C-4 = 9 or missing, then DPVOC = missing.

DPCLSHR (How Many Hours Spent In Prison Classes Last Week)

Derived from C-7. If C-7 = 0, then DPCLSHR = 0; if 1< = C-7 < = 19, then DPCLSHR = 1; if 20 < = C-7< = 49, then DPCLSHR = 2; if C-7 > = 50, then DPCLSHR = 3; if C-7 = 98/99, then DPCLSHR = missing.

DOFFENS1 (Offense 1 For Which Inmate Is In Prison)

Derived from C-8A. The text responses of inmate's offenses to C-8A were recoded into 1 of 5 major offense classifications.

DOFFENS2 (Offense 2 For Which Inmate Is In Prison)

Derived from C-8B. The text responses of inmate's offenses to C-8B were recoded into 1 of 5 major offense classifications. If C-8B = missing, DOFFENS2 = 98.

DOFFENS3 (Offense 3 For Which Inmate Is In Prison)

Derived from C-8C. The text responses of inmate's offenses to C-8C were recoded into 1 of 5 major offense classifications. If C-8C = missing, DOFFENS3 = 98.

DOFFENSE (Type Of Offense For Which Inmate Received Longest Sentence)

Derived from C-8 and C-9. DOFFENSE captures the offense for which the inmate received the longest sentence. Each of C-8A through C-8E was first recoded into 5 major offense classifications. If only 1 offense was listed in C-8, then DOFFENSE = C-8A (recoded). If multiple offenses were listed in C-8, then C-9 was used to determine the offense for which the inmate received the longest sentence. If C-9 = 1, then DOFFENSE = C-8A (recoded); if C-9 = 2, then DOFFENSE = C-8B (recoded); if C-9 = 3, then DOFFENSE = C-8C (recoded); if C-9 = 4, then DOFFENSE = C-8D (recoded); if C-9 = 5, then DOFFENSE = C-8E (recoded). If C-9 = 95/98/99, then DOFFENSE = missing.

DCRIMHIS (Previous Criminal History)

Derived from C-11 and C-12. If C-11 = 1 and C-12 = 1, then DCRIMHIS = 4; if C-11 = 1 and C-12 = 2, then DCRIMHIS = 3; if C-11 = 2 and C-12 = 1, then DCRIMHIS = 2; if C-11 = 2/8 and C-12 = 2/8, then DCRIMHIS = 1.

DRELEASE (Expected Date Of Release)

Derived from C-13, C-14, C-15, and C-16. There were two steps for calculating this variable: 1) determining the respondent's expected year of release and 2) subtracting the year the assessment was administered (2004) from the respondent's expected year of release.

Calculating expected year of release:

1. If C-13 = 1, then calculate expected year of release from the year in C-14. If the year in C-14 = 9998, then DRELEASE = missing.

2. If C-13 = 2 or C-15 = 1, then calculate expected year of release from the year in C-16. If the year in C-16 = 9998 or missing, then DRELEASE = missing.

3. If C-13 = 2 and C-15 = 2, then DRELEASE = 2. These are prisoners who did not expect to be released.

4. If C-13 = missing and C-15 = missing, then DRELEASE = missing. These are prisoners who had not been sentenced when the BQ was administered.

For respondents not classified for DRELEASE in steps 1—4, calculate DRELEASE by subtracting 2004 from expected year of release. Recode the difference to either DRELEASE = 1 or DRELEASE = 2.

DLENGTHC (Length Of Sentence: Collapsed)

Derived from C-10, C-13, C-14, C-15, and C-16. DLENGTHC was recoded from a detailed derived variable, DLENGTHD. DLENGTHD was derived as follows: If C-13 = missing and C-15 = missing, then DLENGTHD = 999999999 (Not sentenced yet). Otherwise, there were two steps for calculating this variable: 1) determining the respondent's expected month and year of release and 2) subtracting the respondent's date of admission to prison from the expected date of release and recoding the date into months.

Note: If the month in C-10/C-14/C-16 equals 98 or missing while the year in C-10/C-14/C-16 is non-missing (i.e., unequal to missing or 9998), set June as the month for the month variables.

Calculating expected month and year of release:

If C-13 = 1, then calculate expected month and year of release from C-14.

If C-13 = 2 or C-15 = 1, then calculate expected month and year of release from C-16.

Once expected month and year of release were calculated:

Use C-10 to get the month and year of admission. Subtract the date of admission from the respondent's expected date of release and recode the date into months.

Note: If the year in C-14/C-16 = missing/9998 or the year in C-10 = missing/9998, then DLENGTHD = missing. If expected date of release is earlier than date of admission, then DLENGTHD = missing.

If C-13 = 2 and C-15 = 2, then DLENGTHD = 99999999997 (Do not expect to be released).

If C-13 = 8/9 and C-15 = 8/9, then DLENGTHD = missing.

DLENGTHC was then recoded from DLENGTHD as follows: If DLENGTHD = 99999999997, then DLENGTHC = 3; if DLENGTHD = 99999999999, then DLENGTHC = missing. Otherwise, if 0< = DLENGTHD< = 60 then DLENGTHC = 1; if 61< = DLENGTHD< = 120 then DLENGTHC = 2; if DLENGTHD> = 121 then DLENGTHC = 3.

DPJOBHR (How Many Hours Worked At Current Job In Prison In The Last Week)

Derived from D-4. If D-4 = 0, then DPJOBHR = 1; if 1 < = D-4 < = 9, then DPJOBHR = 2; if 10 < = D-4< = 19, then DPJOBHR = 3; if 20 < = D-4< = 29, then DPJOBHR = 4; if D-4 > = 30, then DPJOBHR = 5; if D-4 = missing, then DPJOBHR = 6; if D-4 = 98/99, then DPJOBHR = missing.

DPBQ1615 (Worked Full Time Or Not In The Past Three Years While Not In Prison)

Derived from D-11. If D-11 = 1, then DPBQ1615 = 1; if D-11 = 2/3, then DPBQ1615 = 2; if D-11 = missing, then DPBQ1615 = 98.

DPLIBACS (Length Of Time To Access Prison Library)

Derived from E-4, E-5, E-6, and E-7. If E-4 = 1 then DPLIBACS = 1. Otherwise, if E-5 = 2 or E-6 = 2, then DPLIBACS = 5. Otherwise, use E-7 to code DPLIBACS as follows: If E-7 = 8/9, then DPLIBACS = missing; if E-7 = missing, then DPLIBACS = 98; otherwise, DPLIBACS = E-7.