Chapter 9

SCALING AND PROFICIENCY ESTIMATES

Kentaro Yamamoto, Educational Testing Service

The National Adult Literacy Survey results are reported on the same three proficiency scales—prose, document, and quantitative—used for the NAEP 1985 young adult literacy assessment. This chapter describes the models and procedures used to scale the National Adult Literacy Survey results, to estimate respondents' proficiencies, and to conduct statistical analyses.

9.1 SCALING

The National Adult Literacy Survey gathered descriptive and proficiency information on 26,091 sampled respondents through a background questionnaire and a series of assessment booklets containing prose, document, and quantitative literacy tasks. Respondents were sampled using a four-stage stratified sampling method, as described in Chapter 2. In addition to the national sample, several other samples of respondents were surveyed using the same or similar instruments and mode of administration. Eleven states chose to participate in the concurrent State Adult Literacy Survey, each of which surveyed a sample of approximately 1,000 adults: California, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas, and Washington. Florida also surveyed approximately 1,000 adults, but at a later date. These supplementary samples allow results to be reported for these individual states; such information would not be possible if only the state's portion of the national sample were available for analysis. Another supplementary sample included 1,147 respondents incarcerated in 80 state and federal prisons.

All but 1,000 survey respondents who were living in households (that is, who were not in prison) received a monetary incentive of $20 for their participation. Previous studies on the use of incentive payments have found that the absence of an incentive lowers response rates, especially among respondents whose literacy proficiency is low. A response incentive payment of $15 was used in the 1985 young adult literacy assessment. For this reason, the payment of an incentive to the National Adult Literacy Survey respondents maintained comparability. At the request of the Office of Management and Budget, an experimental sample of 1,000 respondents did not receive any incentive, monetary or otherwise, in order to explore further the effects of incentives on the survey results. The results for this non-incentive sample were not included in the National Adult Literacy Survey reports, and are not included in this chapter.

Survey participants spent approximately 20 minutes answering a common set of background questions concerning their demographic characteristics, educational experiences, labor market experiences, and literacy-related activities. Responses to these background questions serve two major purposes. First, they provide a way to summarize the survey results using an array of descriptive variables, such as sex, age, educational attainment, and country of birth. Second, they increase the accuracy of the proficiency estimates for various subpopulations, as described later in this chapter.

The respondents spent the remainder of their time, approximately 45 minutes, completing a booklet of literacy tasks, measuring their prose, document, and quantitative skills. The assessment tasks administered in the National Adult Literacy Survey were created based on a definition of literacy drafted by a panel of experts in the field (see Chapter 4). Most of the cognitive tasks included in the assessment were open-ended or constructed-response questions that required respondents to provide a written answer. A small number of multiple-choice tasks were carried over from the earlier literacy surveys, making it possible to measure trends in performance and to compare the results from different assessments.

A large number of tasks had to be administered in the National Adult Literacy Survey to ensure that the survey would provide the broadest possible coverage of the literacy domains specified. Yet, no individual could be expected to respond to the entire set of 166 simulation tasks (tasks that simulate the demands that adults encounter when they interact with printed materials on a daily basis). Accordingly, the survey was designed using a variant of matrix sampling to give each participant a subset of the total pool of literacy tasks, while at the same time ensuring that each of the 166 tasks was administered to a nationally representative sample of adults.

Respondents' literacy proficiencies are estimated based on their performance on the cognitive tasks administered in the assessment. Unlike multiple-choice questions, which are commonly used in large-scale surveys, open-ended tasks such as those used in the National Adult Literacy Survey elicit a large variety of responses. Verbatim responses must be grouped in some way in order to summarize the performance results. Responses to the open-end tasks of the National Adult Literacy survey were classified into four categories: correct, incorrect, omitted, and not presented.

Since the National Adult Literacy Survey used a variant of matrix sampling and different respondents received different sets of tasks, it would be inappropriate to use any statistic based on the number of correct responses for reporting results, such as the proportion of tasks answered correctly. Differences in total scores (or statistics based on them) between respondents who took a different set of tasks may be caused by differences in respondents' abilities, differences in difficulty between the two sets of tasks, or both. Unless one makes very strong assumptions—for example, that the two sets of tasks are perfectly parallel—the performance of the two groups assessed in a matrix sampling arrangement cannot be directly compared using total score statistics. Moreover, task-by-task reporting ignores the similarities of subgroup comparisons that are common across tasks. Finally, using the average percentage of tasks answered correctly to estimate the proficiency means of examinees in a given subpopulation does not provide any other information about the distribution of skills within that subpopulation.

These limitations of conventional scoring methods can be overcome by using item response theory. When several tasks require similar skills, the response patterns should have some regularity. This regularity can be used to characterize both respondents and tasks in terms of a common scale, even when all respondents do not receive identical sets of tasks in their booklets. In this way, it becomes possible to discuss distributions of performance in a population, or subpopulation, and to estimate the relationships between proficiency and background variables.

The methods and procedures used to analyze the National Adult Literacy Survey results were carefully designed to capture most of the dominant data characteristics. Nevertheless, whatever procedure is used to aggregate data, a certain amount of information is lost when it does not fit the statistical model for proficiency estimates. The data that do not fit must be regarded as inessential to the analyses.

The design of the 1985 NAEP young adult literacy assessment established four proficiency domains—prose, document, quantitative, and reading. For the 1992 National Adult Literacy Survey, scaling was carried out separately for three of these four domains. The 1985 reading scale was dropped from the analyses because what the NAEP reading scale measures had changed in the intervening years. Use of the 1985 block of NAEP reading tasks would no longer be useful for comparisons to the 1992 NAEP reading assessment. The 1992 NAEP reading assessment had changed its block design to 25 minute reading blocks that would not fit the 15-minute block structure of the 1992 National Adult Literacy Survey. Accordingly, the three scales analyzed for the National Adult Literacy Survey were prose literacy, document literacy, and quantitative literacy, but not NAEP reading. By creating a separate scale for each of these domains, it remains possible to explore potential differences in subpopulation performance across these domains. Chapter 12 of this report discusses the rationale for using three distinct scales and examines the correlations among them.

9.2 SCALING METHODOLOGY

This section reviews the scaling model employed in the analyses of the National Adult Literacy Survey data and describes the plausible values methodology used for proficiency estimation.

9.2.1 The Scaling Model

The scaling model used for the National Adult Literacy Survey is the three-parameter logistic (3PL) model from item response theory (Birnbaum, 1968; Lord, 1980). It is a mathematical model for estimating the probability that a particular person will respond correctly to a particular task from a single domain of tasks. This probability is given as a function of a parameter characterizing the proficiency of a given person, and three parameters characterizing the properties of a given task. The following three-parameter logistic item response theory model was employed in the National Adult Literacy Survey:

$$P(x_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i(\theta_j - b_i)}} \qquad (1)$$

where

x_{ij} is the response of person j to task i, 1 if correct and 0 if incorrect;

θ_j is the proficiency of person j (note that a person with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of task i, characterizing its sensitivity to proficiency;

b_i is its locator parameter, characterizing its difficulty; and

c_i is its lower asymptote parameter estimated only for the multiple-choice tasks, reflecting possibly non-zero chances of correct response, even for persons with very low proficiencies; for open-ended tasks, c was fixed at zero.

Note that this is a monotonically increasing function with respect to θ; that is, the conditional probability of a correct response increases as the value of θ increases. In addition, a linear indeterminacy exists with respect to the values of θ_j, a_i, and b_i for a scale defined under the three-parameter model. In other words, for an arbitrary linear transformation of θ, say $\theta^* = M\theta + X$, the corresponding transformations $a^*_i = a_i/M$ and $b^*_i = Mb_i + X$ give:

$$P(x_{ij} = 1 | \theta^*_j, a^*_i, b^*_i, c^*_i) = P(x_{ij} = 1 | \theta_j, a_i, b_i, c_i) \qquad (2)$$

Linear transformation of the scales was used to link the National Adult Literacy Survey scales to the 1985 young adult literacy assessment scales for gain purposes. The scale indeterminacy was resolved by setting an origin and unit size of θ to the reported scale means and standard deviations from 1985 young adult literacy assessment.

The main assumption of item response theory is conditional independence. In other words, item response probabilities depend only on θ (a measure of proficiency) and the specified item parameters, as opposed to depending on any demographic characteristics of examinees, or on any other items presented together in a test, or on the survey administration conditions. Controlling for θ, the probability of a correct response on one item is unrelated to the probability of a correct response on another given θ. This allows one to formulate the following joint probability of a particular response pattern *x* across a set of n items.

$$\mathrm{P}(\mathbf{x}|\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}_{\mathrm{i}}(\theta)^{\mathrm{x}_{\mathrm{i}}} (1 - \mathrm{P}_{\mathrm{i}}(\theta))^{1-\mathrm{x}_{\mathrm{i}}} \qquad (3)$$

By replacing the hypothetical response pattern with the real scored data, one can view the above function as a likelihood function that is to be maximized with a given set of item parameters. These item parameters were treated as known for the subsequent analyses.

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

Testing the assumptions of the item response theory model, especially the assumption of conditional independence, is a critical part of the data analyses. Serious violation of the conditional independence assumption would undermine the accuracy and integrity of the results. Thus, while the item parameters were being estimated, empirical distribution of percentages correct conditional on θ and the item parameters were monitored across the adult sample of individuals 16 to 65 and the sample of adults over 65. For a few tasks, the percentages of correct responses obtained by the older sample were quite different from those obtained by the younger sample, and these tasks were dropped from the National Adult Literacy Survey analyses.

9.2.2 Design for Linking the 1992 Scales to the 1985 Scales

As previously noted, the prose, document, and quantitative literacy results for the National Adult Literacy Survey are reported on scales that were established in the 1985 young adult literacy assessment. Eighty-five (51 percent) of the tasks administered in the 1992 National Adult Literacy Survey were originally administered in 1985. The linkage between the scales from the two surveys is based on these tasks. In addition, 81 new tasks were developed for the National Adult Literacy Survey. A total of 166 tasks were administered in the 1992 survey. The composition of the National Adult Literacy Survey item pool is presented in Table 9-1.

Table 9-1. Composition of item pool for the National Adult Literacy Survey

Literacy scale	Number of tasks common to 1985 and 1992	Number of tasks in 1992 only	Total in 1992
Prose	14	27	41
Document	56	26	81
Quantitative	15	28	43
Total	85	81	166

A unidimensional item response theory model like the three-parameter logistic one employed in this study assumes that performance on all the items in a domain can, for the most part, be accounted for by a single (unobservable) proficiency variable. Subsequent linking and scaling analyses treated each scale separately—that is, a unique proficiency was assumed for each scale. As a result, the linking of corresponding scales was carried out for each scale separately. The three steps used to link the 1985 and 1992 scales are listed below.

1. Establish provisional item response theory scales through common item parameter calibrations based on a pooling of the 1992 and 1985 tasks.
2. Estimate the distribution of proficiencies on the provisional item response theory scales using plausible values.
3. Align the 1992 National Adult Literacy Survey scales to the 1985 scales by a linear transformation based on the common proficiency distribution of the 1985 sample.

9.2.3 Item Parameter Estimation

Identical item calibration procedures, described here in detail, were carried out separately for each of the three literacy scales. Using a modified version of Mislevy and Bock’s (1982) BILOG computer program, the three-parameter logistic item response theory model was fit to each task (but with lower asymptote parameters fixed at zero for open-ended tasks) using sample weights.

The cognitive tasks administered in the 1985 young adult literacy assessment were used for several assessments and surveys, including the National Adult Literacy Survey, surveys in Oregon and Mississippi, the 1989-90 survey of job-seekers conducted for the U.S. Department of Labor, and a second Department of Labor assessment. In total, more than 40,000 individuals have responded to either the entire set or a subset of the 1985 young adult literacy assessment tasks. To obtain stable item parameter estimates and simplify scale linking procedures, the data accumulated from all surveys were included in a calibration sample. The current method of parameter calibration in effect puts all available survey results on a single provisional common scale. Only linear indeterminacy needed to be resolved in order to align the provisional scale to the reporting scale.

Sample weights were used during item calibration. It is known that different subpopulation distributions occur within different assessment samples. Such variations may arise because of differences in the characteristics of the target populations, the sampling design, or the randomness of sampling. For example, oversampling of racial/ethnic minority populations is often necessary to ensure a certain degree of accuracy in estimating group proficiencies. In such cases, the unweighted sample would not represent the targeted population correctly. Post-stratified weights take into account the sampling design, such as oversampling as well as the randomness of real data. By applying post-stratified weights, vital characteristics of the sample can be closely matched to the characteristics of the population. During calibration, the fit of item parameters is maximized in reference to the proficiency distribution of the calibration sample. When item parameters are being estimated, it is ideal to match the proficiency distribution of the calibration sample as closely as possible to that of the population. It is more critical when item calibration is done on the combined proficiency distribution of multiple assessment samples with great differences in proficiency distributions, such as the National Adult Literacy Survey. It was not as critical for the analysis of the 1985 young adult literacy assessment results because the young adult item parameters were estimated based on one sample.

To obtain unbiased parameter estimates, proficiency distributions for the separate assessment samples were estimated during calibration. In addition to the samples from the previous assessments, certain groups in the National Adult Literacy Survey respondents received separate proficiency distributions; those included adults age 16 to 64, those age 65 and older, prisoners, and respondents who received no monetary incentive for participating in the survey. It is known that the samples for each assessment came from somewhat different populations with different characteristics. In addition, the number of tasks administered varied in each assessment. The calibration procedure should take into account the possibility of systematic interaction of samples and tasks to generate unbiased estimates of sample distributions and item parameters. For that reason, a normal distribution with a unique mean and variance for each assessment population was estimated concurrently with item parameters. Estimated item parameters for each literacy scale are presented in Tables 9-2p, 9-2d, and 9-2q.

Model fit was evaluated at the task level by examining BILOG likelihood ratio chi-square statistics for each survey sample.[1] The fit was also evaluated by inspecting residuals from fitted item response curves. A typical plot is shown in Exhibit 9-1.

In Exhibit 9-1, the horizontal axis represents the provisional proficiency scale derived directly from the calibration procedure. The provisional scale is in standard units, without transformation to the 0

[1] The sampling distributions are probably not strictly χ^2 with the indicated degrees of freedom. Therefore, they were used as descriptive indices of relative model fit rather than as a statistical test of fit.

to 500 scale used for other purposes. The smooth curved line is the fitted three-parameter logistic item response curve. Each calibration sample is represented by a unique plot symbol. The five plot symbols represent the (approximate) expected proportions of correct responses at various points along the scale. The size of the plot symbols is proportional to the information available in the calibration data in that region of the scale. In general, the fit of the model was quite good. For some tasks, there was evidence that the estimated parameters did not fit certain assessment samples as well as other samples; however, this pattern was not consistently apparent for any one sample. Five tasks were dropped from calibration due to a lack of fit.

Table 9-2p. Prose literacy item descriptions and parameters for the National Adult Literacy Survey

Number	Description	A	SE(A)	B	SE(B)	C	SE(C)
AB21101	Swimmer: Underline sentence telling what Chanin ate	1.125	0.042	-1.901	0.048	0.000	0.000
AB21201	Swimmer: Age Chanin began to swim competitively	1.070	0.029	-1.124	0.027	0.000	0.000
AB30501	Technology: Underline sentence explaining action	0.590	0.015	0.593	0.022	0.000	0.000
AB30601	Technology: Orally explain info from article	0.915	0.023	0.347	0.018	0.000	0.000
AB31201	Dickinson: Describe what is expressed in poem	0.725	0.018	0.691	0.020	0.000	0.000
AB40901	Korean Jet: Give argument made in article	0.826	0.018	0.165	0.017	0.000	0.000
AB41001	Declaration: Describe what poem is about	0.622	0.020	-1.433	0.053	0.000	0.000
AB50101	Panel: Find information from article	0.466	0.016	2.112	0.057	0.000	0.000
AB50201	Panel: Determine surprising future headline	1.160	0.036	0.861	0.017	0.196	0.000
AB60201	Make out check: Write letter explaining bill error	1.240	0.027	-0.440	0.015	0.000	0.000
AB60601	Economic index: Underline sent. Explaining action	0.808	0.019	-0.319	0.021	0.000	0.000
AB70401	Almanac vitamins: List correct info from almanac	0.705	0.018	-0.765	0.029	0.000	0.000
AB71001	Instruction to return appliance: Indicate best note	1.377	0.042	-0.305	0.020	0.266	0.000
AB71101	Explain difference between 2 types of benefits	0.782	0.021	0.482	0.021	0.000	0.000
NC00301	"My Dream:" Find country in short story	0.892	0.034	-3.228	0.090	0.000	0.000
NC00401	"My Dream:" Underline sentence explaining action	0.765	0.016	-1.935	0.034	0.000	0.000
N010101	Marketing: List two facts	0.868	0.025	0.607	0.022	0.000	0.000
N010201	Marketing: Underline sentence explaining action	1.059	0.031	-0.402	0.022	0.000	0.000
N010301	Marketing: Give purpose of event	0.786	0.031	2.138	0.053	0.000	0.000
N080101	SSI: Mark correct information in article	1.328	0.051	-1.447	0.036	0.000	0.000
N080201	SSI: What must an SSI user accept if offered?	1.516	0.043	-0.389	0.017	0.000	0.000
N080301	SSI: What is most you can make to receive SSI?	0.618	0.021	0.486	0.028	0.000	0.000
N090601	Face off: What group will mandate safe cars?	1.878	0.064	-0.748	0.018	0.000	0.000
N090701	Face off: Find correct information in article	1.804	0.060	-0.699	0.018	0.000	0.000
N090801	Contrast views on fuel-efficiency vs. size of car	1.239	0.037	1.091	0.020	0.000	0.000
N100101	"Growing Up:" Find first buyer's name	1.466	0.052	-1.146	0.027	0.000	0.000
N100201	"Growing Up:" Determine correct day of delivery	1.297	0.037	-0.345	0.018	0.000	0.000
N100301	"Growing Up:" What reason given to stop selling?	1.187	0.034	-0.343	0.020	0.000	0.000
N100401	"Growing Up:" Compare approaches to selling mags	0.841	0.027	1.236	0.029	0.000	0.000
N110101	Blood pressure: Why difficult to know if high	0.988	0.032	-0.971	0.032	0.000	0.000
N110401	Jury: Length of time served by a juror	0.770	0.024	-0.191	0.027	0.000	0.000
N110501	Jury: Underline sentence explaining action	0.939	0.030	-0.730	0.030	0.000	0.000
N110601	Two challenges attorneys use to jurors	1.044	0.039	1.954	0.038	0.000	0.000
N120301	Ida Chen: What experience turned Ida toward law?	1.074	0.030	0.141	0.019	0.000	0.000
N120401	Two things Chen did to resolve discrimination conflicts	1.162	0.032	0.229	0.017	0.000	0.000
N120501	Ida Chen: Interpret phrase from article	0.926	0.037	2.107	0.048	0.000	0.000
N120901	Susan Butcher: Find number of wins of sled race	0.888	0.044	-2.061	0.080	0.000	0.000
N130201	Fueled: Determine phrase meaning	1.089	0.030	0.315	0.018	0.000	0.000
N130301	Fueled: Give diff and similarity between events	0.978	0.030	1.213	0.025	0.000	0.000
N130401	Fueled: Give suggestion about good value change	1.576	0.045	0.978	0.016	0.000	0.000
N130801	Cost to raise child: Find information from article	0.735	0.027	-1.012	0.043	0.000	0.000

Table 9-2d. Document literacy item descriptions and parameters for the National Adult Literacy Survey

Item#	Description	A	SE(A)	B	SE(B)	C	SE(C)
SCOR100	Social Security card: Sign name on line	0.504	0.025	-4.803	0.248	0.000	0.000
SCOR300	Driver's license: Locate expiration date	0.917	0.025	-2.525	0.058	0.000	0.000
AB20101	Energy graph: Find answer for given conditions (1)	1.153	0.045	-0.193	0.054	0.228	0.030
AB20201	Energy graph: Find answer for given conditions (2)	0.935	0.030	-0.023	0.045	0.096	0.023
AB20301	Energy: Yr 2000 source percent power larger than 71	1.089	0.036	0.684	0.031	0.142	0.015
AB20401	Yellow pages: Find a list of stores	0.478	0.019	-0.467	0.111	0.144	0.036
AB20501	Yellow pages: Find telephone number of given place	0.414	0.017	-0.771	0.111	0.088	0.031
AB20601	Yellow pages: Find place open Saturday	1.077	0.034	-0.143	0.041	0.105	0.023
AB20701	Bus schd: Take correct bus for given condition (1)	0.521	0.024	0.293	0.106	0.130	0.035
AB20801	Bus schd: Take correct bus for given condition (2)	1.282	0.044	0.901	0.024	0.144	0.012
AB20901	Bus schd: After 2:35, how long til Flint&Acad bus	1.168	0.032	1.520	0.021	0.162	0.008
AB21001	Bus schd: Take correct bus for given condition (4)	0.730	0.031	0.520	0.066	0.144	0.026
AB21501	With graph, predict sales for spring 1985	0.799	0.024	-0.571	0.038	0.000	0.000
AB30101	Street map: Locate intersection	0.953	0.027	-0.956	0.036	0.000	0.000
AB30301	Sign out sheet: Respond to call about resident	0.904	0.025	-0.844	0.034	0.000	0.000
AB30401	Sign out sheet: Respond to call about resident (2)	0.665	0.017	-0.089	0.028	0.000	0.000
AB30701	Major medical:locate Eligibility from table	0.960	0.026	-0.702	0.030	0.000	0.000
AB30801	Almanac: Find page containing chart for given info	0.704	0.017	0.929	0.019	0.000	0.000
AB30901	Almanac: Determine pattern in exports across years	0.299	0.013	0.000	0.056	0.000	0.000
AB31001	Abrasive guide: Type of sandpaper for sealing	0.831	0.019	0.285	0.020	0.000	0.000
AB31101	Abrasive gd: Can product be used in given case?	0.761	0.020	-0.256	0.028	0.000	0.000
AB31301	Facts about fire: Mark information in article	0.721	0.024	-1.170	0.055	0.000	0.000
AB40101	School registration: Mark correct age information	0.820	0.024	-1.063	0.041	0.000	0.000
AB40401	Almanac: Find page containing chart for given info	1.108	0.023	0.717	0.013	0.000	0.000
AB50401	Catalog order: Order product one	0.772	0.022	-0.882	0.039	0.000	0.000
AB50402	Catalog order: Order product two	0.771	0.018	0.396	0.019	0.000	0.000
AB50501	Telephone bill: Mark information on bill	0.359	0.014	-0.511	0.060	0.000	0.000
AB50601	Almanac football: Locate page of info in almanac	1.001	0.023	-0.083	0.020	0.000	0.000
AB50701	Almanac football: Explain why an award is given	1.182	0.029	-0.373	0.022	0.000	0.000
AB50801	Wage & tax statement: What is current net pay?	0.733	0.025	-1.365	0.060	0.000	0.000
AB50901	Wage & tax statement: What is yr-to-date gross pay	0.884	0.022	-0.199	0.025	0.000	0.000
AB60101	Make out check: Enter correct date on check	1.254	0.031	-0.497	0.021	0.000	0.000
AB60102	Make out check: Paid to the correct place	1.408	0.035	-0.425	0.018	0.000	0.000
AB60103	Make out check: Enter correct amount in numbers	0.993	0.026	-0.674	0.028	0.000	0.000
AB60104	Make out check: Enter correct amount written out	1.537	0.040	-0.524	0.018	0.000	0.000
AB60301	Phone message: Write correct name of caller	1.454	0.054	-1.283	0.036	0.000	0.000
AB60302	Phone message: Write correct number of caller	1.068	0.038	-1.434	0.048	0.000	0.000
AB60303	Phone message: Mark “please call” box	0.903	0.024	-0.680	0.030	0.000	0.000
AB60304	Phone message: Write out correct message	0.895	0.019	0.461	0.017	0.000	0.000
AB60305	Phone message: Write who took the message	0.640	0.017	-0.220	0.030	0.000	0.000
AB60306	Phone message: Write whom message is for	0.947	0.027	-0.867	0.033	0.000	0.000
AB60501	Petroleum graph: Label axes of graph	1.102	0.024	1.937	0.019	0.000	0.000
AB60502	Petroleum graph: Complete graph including axes	1.081	0.023	0.782	0.014	0.000	0.000
AB60701	Nurses’ convention: Who would be asked questions	1.179	0.045	-1.295	0.047	0.000	0.000

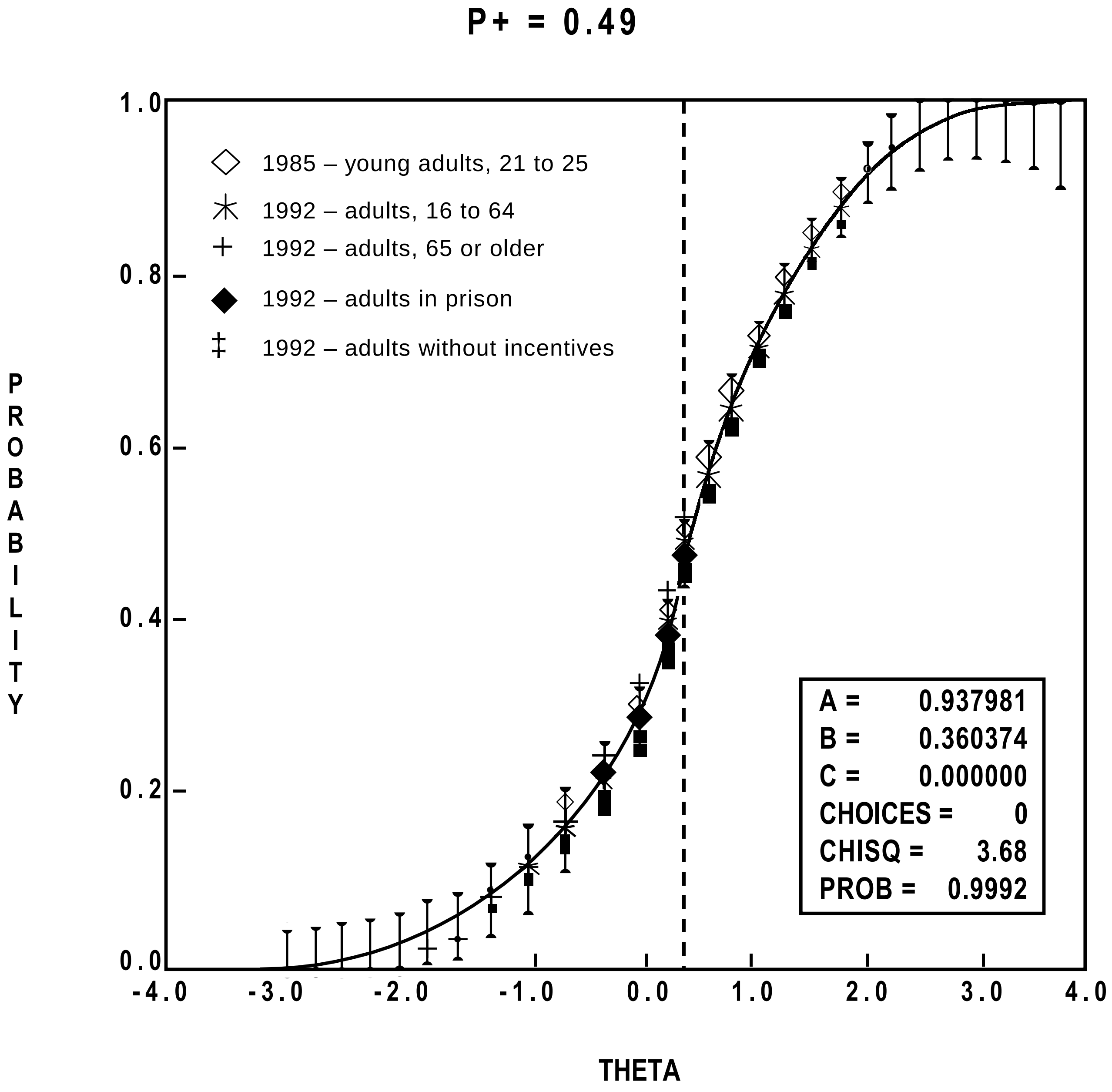
Table 9-2d. Document literacy item descriptions and parameters for the National Adult Literacy Survey — Continued

Item#	Description	A	SE(A)	B	SE(B)	C	SE(C)
AB60801	Nurses' convention: Write correct day of program	1.016	0.042	-1.539	0.063	0.000	0.000
AB60802	Nurses' convention: What is date of program?	1.231	0.058	-1.620	0.064	0.000	0.000
AB60803	Nurses' convention: What is time of program?	1.438	0.076	-1.649	0.063	0.000	0.000
AB61001	Nurses' convention: Write correct place for tables	0.766	0.030	-1.454	0.069	0.000	0.000
AB70104	Job application: Complete personal information	0.542	0.024	-2.337	0.119	0.000	0.000
AB70301	Almanac vitamins: Locate list of info in almanac	0.753	0.018	-0.134	0.025	0.000	0.000
AB70701	Follow directions on map: Give correct location	0.799	0.019	-0.126	0.024	0.000	0.000
AB70801	Classified: Match list with coupons	1.142	0.033	-0.880	0.030	0.000	0.000
AB70901	Checking deposit: Enter correct date	0.990	0.030	-1.088	0.039	0.000	0.000
AB70902	Checking deposit: Enter correct cash amount	0.858	0.021	-0.303	0.025	0.000	0.000
AB70903	Checking deposit: Enter correct amount of check	1.266	0.038	-0.921	0.029	0.000	0.000
AB71201	Mark correct movie from given information	0.939	0.041	-1.801	0.077	0.000	0.000
N010401	Vehicle chart: Find correct information	0.902	0.038	-1.340	0.062	0.000	0.000
N010801	Trend chart: Mark information on chart	0.807	0.028	-0.463	0.038	0.000	0.000
N010901	Trend chart: Put information on chart	0.720	0.024	1.702	0.032	0.000	0.000
N011001	Trend chart: Determine least # of points needed	0.645	0.022	0.260	0.032	0.000	0.000
N080601	Bus schedule: Take correct bus for given condition	1.039	0.029	0.505	0.020	0.000	0.000
N080701	Bus schedule: Mark map correctly for given info	1.094	0.034	-0.312	0.027	0.000	0.000
N080801	Auto maintenance form: Enter information given (1)	0.763	0.023	0.569	0.025	0.000	0.000
N080802	Auto maintenance form: Enter given information	1.357	0.048	-0.683	0.029	0.000	0.000
N090301	Essence: Determine page certain article begins on	1.123	0.048	-1.224	0.051	0.000	0.000
N090401	Essence: Determine topic of given article	0.987	0.033	-0.448	0.032	0.000	0.000
N090501	Essence: Determine topic of section of magazine	0.671	0.024	-0.301	0.040	0.000	0.000
N100501	Opinions table: Mark sentence explaining action	1.038	0.029	0.486	0.020	0.000	0.000
N100601	Opinions table: Find correct group for given info	1.134	0.032	1.284	0.019	0.000	0.000
N100701	Summarize views of parents & teachers	1.127	0.034	2.300	0.032	0.000	0.000
N110301	Certified mail rec't: Enter name and address	0.811	0.029	-0.742	0.045	0.000	0.000
N110302	Certified mail rec't: Enter postage and fee	0.714	0.028	-1.025	0.059	0.000	0.000
N110701	Credit card table: Find correct bank	0.469	0.020	0.125	0.047	0.000	0.000
N110901	Credit card table: Give 2 differences	0.829	0.031	1.882	0.032	0.000	0.000
N120101	Campus map: Mark map for given info	0.985	0.036	-0.801	0.040	0.000	0.000
N120201	Campus map: Find correct room for given dean	0.842	0.028	-0.403	0.035	0.000	0.000
N120601	Middle class: Find projected percent	0.795	0.037	-1.488	0.077	0.000	0.000
N130101	S.S. card application: Identify and enter info(1)	1.619	0.049	-0.095	0.017	0.000	0.000
N130102	S.S. card application: Identify and enter info(3)	1.270	0.043	-0.544	0.028	0.000	0.000
N130103	S.S. card application: Identify and enter info(2)	2.105	0.071	-0.290	0.016	0.000	0.000
N130104	S.S. card application: Identify and enter info(4)	2.159	0.069	-0.111	0.014	0.000	0.000

Table 9-2q. Quantitative literacy item descriptions and parameters for the National Adult Literacy Survey

Number	Description	A	SE(A)	B	SE(B)	C	SE(C)
AB40201	Unit price: Estimate cost/oz of peanut butter	0.818	0.019	0.455	0.017	0.000	0.000
AB40301	Unit price: Mark economical brand	0.815	0.034	0.216	0.029	0.447	0.000
AB40501	Airline schedule: plan travel arrangements (1)	0.909	0.020	0.005	0.016	0.000	0.000
AB40601	Airline schedule: plan travel arrangements (2)	0.952	0.021	-0.522	0.018	0.000	0.000
AB40701	Check ledger: Complete ledger (1)	1.597	0.034	-0.500	0.013	0.000	0.000
AB40702	Check ledger: Complete ledger (2)	1.936	0.042	-0.344	0.010	0.000	0.000
AB40703	Check ledger: Complete ledger (3)	1.873	0.040	-0.331	0.011	0.000	0.000
AB40704	Check ledger: Complete ledger (4)	1.970	0.042	-0.294	0.010	0.000	0.000
AB50301	Interest charges: Orally explain computation	0.601	0.020	1.522	0.043	0.000	0.000
AB50403	Catalog order: Order product three	0.609	0.016	0.600	0.023	0.000	0.000
AB50404	Catalog order: Shipping, handling, and total	0.968	0.023	-0.951	0.022	0.000	0.000
AB60901	Nurses Convention: Write number of seats needed	0.504	0.015	-0.355	0.031	0.000	0.000
AB70501	Lunch: Determine correct change using info in menu	0.893	0.019	0.090	0.016	0.000	0.000
AB70601	Lunch: Determine 10% tip using given info	0.872	0.019	0.384	0.016	0.000	0.000
AB70904	Checking deposit: Total bank deposit entry	0.869	0.029	-1.970	0.049	0.000	0.000
NC00501	Enter total amount of both checks being deposited	0.661	0.017	-2.792	0.060	0.000	0.000
NC00601	Price for Sleuth: how much less than On the Town	0.717	0.013	-1.690	0.028	0.000	0.000
N010501	Vehicle chart: Find sum of percentages	0.851	0.026	-0.768	0.029	0.000	0.000
N010601	Vehicle chart: Describe solution to percent problem	1.121	0.032	0.717	0.019	0.000	0.000
N010701	Vehicle chart: Find magnitude of difference	1.033	0.029	0.411	0.019	0.000	0.000
N011101	Gas gauge: Use info to answer question-show calcs	1.034	0.030	0.195	0.019	0.000	0.000
N080401	SSI: Calculate yrly amount for couple w/ basic ssi	0.696	0.022	0.520	0.026	0.000	0.000
N080501	Minutes from student union to 17th & Main	0.757	0.023	-0.247	0.025	0.000	0.000
N080901	Auto maintenance form: Calculate miles per gallon	0.850	0.027	0.856	0.026	0.000	0.000
N081001	Rank juices by expense and give reasons	0.732	0.023	0.122	0.025	0.000	0.000
N090101	Get discount if oil bill paid in 10 days	1.346	0.037	-0.018	0.016	0.000	0.000
N090201	Get net total owed after deduction	1.677	0.047	-0.349	0.015	0.000	0.000
N090901	Carpet ad: Get diff in reg and sale price	0.789	0.028	-1.003	0.040	0.000	0.000
N091001	Carpet ad: Get total cost to carpet room	0.634	0.026	1.371	0.045	0.000	0.000
N100801	Salt River: Determine difference in costs	0.647	0.027	-1.737	0.068	0.000	0.000
N100901	Salt River: Determine miles between stops	0.622	0.022	-0.263	0.032	0.000	0.000
N101001	Salt River: Determine hours between points	0.943	0.031	-0.837	0.031	0.000	0.000
N110201	Blood pressure: Calculate death rate from info	1.033	0.030	0.740	0.021	0.000	0.000
N110303	Certified mail rec’t: Calculate postage and fees	0.789	0.031	-1.730	0.056	0.000	0.000
N110801	Credit card table: Determine difference in rates	0.881	0.029	-0.494	0.029	0.000	0.000
N120701	Calc percent diff black & white middle class-1980	0.909	0.029	-0.845	0.029	0.000	0.000
N120801	Middle class: Find difference in magnitude of pct	1.013	0.030	0.830	0.022	0.000	0.000
N121001	Calc miles/day Butcher went in this year's race	1.017	0.031	0.217	0.020	0.000	0.000
N121101	Susan Butcher: Calc diff in times for completion	0.959	0.035	1.517	0.035	0.000	0.000
N130501	Rec room: Calculate feet of molding needed	0.655	0.023	0.819	0.032	0.000	0.000
N130601	Rec room: Calculate number of wall panels needed	1.111	0.031	-0.184	0.019	0.000	0.000
N130701	Rec room: Describe solution of calculation needed	0.845	0.034	1.962	0.052	0.000	0.000
N130901	Raise child: Calc money needed to raise child	0.945	0.030	0.499	0.022	0.000	0.000

Exhibit 9-1. Item response curve for a task included in both the 1985 Young Adult Literacy Assessment and the 1992 National Adult Literacy Survey



9.3 PROFICIENCY ESTIMATION USING PLAUSIBLE VALUES

9.3.1 Generating Proficiency Scores

The purpose of most cognitive skills testing is to accurately assess individual performance for the purposes of diagnosis, selection, or placement. Regardless of which measurement model is being used, classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, achievement tests containing more than 70 items are common. Since the uncertainty associated with each θ is negligible, the distribution of θ or the joint distribution of θ with other variables can be approximated using individual θ's.

When analyzing the distribution of proficiencies for a group, however, more efficient estimates can be obtained from a sampling design like the one used in the National Adult Literacy Survey. The survey solicits relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses are summed for all respondents. The advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. Uncertainty associated with individual θ estimates is too large to be ignored. Point estimates of proficiency that are, in some sense, optimal for each sampled respondent could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, and Beaton, 1987).

Plausible values methodology was developed as a way to estimate key population features consistently and to approximate others at the level of item response theory procedures. Mislevy (1991) provides a detailed review of plausible values methodology. Along with theoretical justifications, Mislevy presents comparisons with standard procedures, discusses biases that arise in some secondary analyses, and offers numerical examples.

The following is a brief overview of the plausible values approach, focusing on its implementation in the 1992 National Adult Literacy Survey analyses.

Let *y* represent the responses of all sampled respondents to background questions and questions on engagement to literacy activities, and let θ represent the scale proficiency values. If θ were known for all sampled examinees, it would be possible to compute a statistic t(θ ,*y*)—such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T.

Because the scaling models are latent variable models, however, θ values are not observed even for sampled respondents. To overcome this problem, we follow Rubin (1987) by considering θ as "missing data" and approximate t(θ,*y*) by its expectation given (x,*y*), the data that actually were observed, as follows:

$$\begin{aligned} \mathrm{t}^{*}(\mathbf{x},\mathbf{y}) &= \mathrm{E}[\mathrm{t}(\theta,\mathbf{y})|\mathbf{x},\mathbf{y}] \\ &= \int \mathrm{t}(\theta,\mathbf{y})p(\theta|\mathbf{x},\mathbf{y})\mathrm{d}\theta \end{aligned} \qquad (4)$$

It is possible to approximate t^{*} using random draws from the conditional distribution of the scale proficiencies given the item responses x_{j}, background variables y_{j}, and model parameters for sampled respondent j. These values are referred to as imputations in the sampling literature, and as plausible values in the National Adult Literacy Survey and in the National Assessment of Educational Progress. The value of θ for any respondent that would enter into the computation of *t* is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed to repeat this process several times so that the uncertainty associated with imputation can be quantified by "multiple imputation." For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects uncertainly due to not observing θ. It should be noted that this variance does not include the variability of sampling from the population.

It cannot be emphasized too strongly that plausible values are not test scores for individuals in the usual sense. Plausible values are only intermediary computations for calculating integrals as shown in the above equation in order to estimate population characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that are in some sense optimal for each respondent (e.g., maximum likelihood estimates, which are consistent estimates of a respondent's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual respondents have distributions that can produce decidedly nonoptimal (inconsistent) estimates of population characteristics (Little and Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

Plausible values for each respondent j are drawn from the multivariate normal conditional distribution $P(\underline{\theta}_j|x_j,y_j,\Gamma,\Sigma)$, where Γ is a matrix of regression coefficients and Σ is a common variance matrix for residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as follows

$$\begin{aligned} P(\theta_j|x_j,y_j,\Gamma,\Sigma) &\propto P(x_j|\theta_j,y_j,\Gamma,\Sigma)P(\theta_j|y_j,\Gamma,\Sigma) \\ &= P(x_j|\theta_j)P(\theta_j|y_j,\Gamma,\Sigma) \end{aligned} \tag{5}$$

where $\underline{\theta}_j$ is a vector of three scale values, $P(x_j \mid \theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\underline{\theta}_j \mid y_j, \Gamma, \Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of background responses and parameters Γ and Σ. Item parameter estimates are fixed and regarded as population values in the computation described in this section. (See Appendix C for Γ (Gamma) values.)

In the National Adult Literacy Survey analyses, a normal multivariate distribution was assumed for $P(\boldsymbol{\theta}_j \mid y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the first approximately principal components of several hundred selected main effects and two-way interactions of the complete vector of background variables. The background variables included sex, ethnicity, Spanish language interview, region of the country, respondent education, parental education, occupation, and reading practices. The complete set of original background variables used in the analyses is listed in Appendix G. Based on the principal component method, components representing 99 percent of the variance present in the data were selected. The included principal components will be referred to as the conditioning variables, and denoted as y^c. The following model was fit to the data:

$$\theta = \Gamma^{c} + \varepsilon \qquad (6)$$

where ε is normally distributed with mean zero and variance Σ. As in a regression analysis, Γ is a matrix each of whose columns is the effects for one scale and Σ is the three-by-three matrix variance of residuals between scales.

Note that in order to be strictly correct for all functions Γ of θ, it is necessary that $p(\theta \mid \mathbf{y})$ be correctly specified for all background variables in the survey. In the National Adult Literacy Survey, principal component scores were generated from background variables. Marginal means and percentile points of θ for these variables can be consistently estimated. Estimates of functions T involving background variables not conditioned in this manner are subject to error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991). Their magnitudes diminish as each respondent provides more cognitive data—that is, responds to a greater number of items. Indications are that the magnitude of these errors is negligible in the National Adult Literacy Survey (e.g., biases in regression coefficients below 5 percent) due to the larger numbers of cognitive tasks presented to each respondent in the survey (on average, 13 tasks per scale). The exception is the sample of respondents who could not or did not proceed beyond the background questions.

These respondents did not attempt the assessment tasks due to an inability to read or write English, a physical disability, a mental disability, or a refusal to participate in the survey. Chapter 8 describes the procedure used to estimate the proficiencies of those with missing responses. If these respondents had been excluded from the survey, the proficiency scores of some subpopulations in the National Adult Literacy Survey would have been severely overestimated, and the picture of the nation’s literacy skills would have been distorted. These respondents possess few literacy skills, and detailed analyses of their proficiencies, not surprisingly, may lead to unstable results.

The basic method for estimating Γ and Σ with the EM procedure was described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean, θ, and variance, Σ, of the posterior distribution. For the multiple scales of the National Adult Literacy Survey, the computer program C-GROUP (Thomas, 1993) was used. The program implemented a method to compute the moments using higher order asymptotic corrections to a normal approximation. Case weights were employed in this step.

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Σ for all sampled respondents with more than four cognitive tasks attempted. First, a value of Γ is drawn from a normal approximation to P(Γ, Σ | x_j, y_j) that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean θ, and variance Σ_j^P of the posterior distribution are computed using the same methods applied in the EM algorithm. In the third step, the θ values are drawn independently from a multivariate normal distribution with mean θ and variance Σ_j^P . These three steps are repeated five times, producing five imputations of θ for each sampled respondent.

For those with an insufficient number of responses, the Γ and Σs described in the previous paragraph were fixed. Hence, all respondents—regardless of the number of tasks attempted—were assigned a set of plausible values for the three scales. The plausible values can then be employed to evaluate an arbitrary function T according to the following five steps:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of θ. Denote the result T_1.

2. In the same manner as in step 1 above, evaluate the sampling variance of T, or Var($T_{1,}$), with respect to respondents' first vectors of plausible values. Denote the result Var_1.

3. Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for u=2,…,5.

4. The best estimate of T obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$T_{.} = \frac{\sum_u T_u}{5} \qquad (7)$$

5. An estimate of the variance of T. is the sum of two components: an estimate of Var(T_u) obtained as in step 4 and the variance among the T_us:

$$\mathrm{Var}(T_{.}) = \frac{\sum_u T_u}{5} + \left(1 + \frac{1}{5}\right) \sum_u (T_u - T_{.})^2 \qquad (8)$$

The first component in Var(T.) reflects uncertainty due to sampling respondents from the population; the second component reflects uncertainty due to the fact that the θs of the sampled respondents are not known precisely, but only indirectly through x and y.

9.3.2 Linking the 1992 Scale to the 1985 Scale

At this point, plausible values are still on the provisional scale and must be transformed to the 1985 scale for comparison. The 1985 scale was established in the following manner. In the 1985 assessment, some of the tasks administered were the same as those included in the NAEP 1984 reading assessment. Relying on the common tasks from the two assessments, the 1985 sample proficiency distribution was placed on the NAEP reading scale, a 0 to 500 metric. The mean and standard deviation of the plausible values for the 1985 samples were estimated to be 296.6 and 49.0, respectively. The mean and standard deviation of the other three scales—prose, document, and quantitative—were also set to these values.

In the 1992 National Adult Literacy Survey, as noted earlier, item parameters from the 1985 young adult literacy assessment were re-estimated using a larger sample and more accurate procedures than were available at the time of the 1985 analysis. These new item parameters are best suited for comparing performance distributions for different samples. However, the new sets of item parameters on the provisional scales and the old transformation constants used to produce the 1985 scales would not necessarily produce identical results for the 1985 sample. Thus, new linear transformation constants for the 1985 sample were found to match the mean and standard deviation of the current plausible value distribution of the 1985 sample based on the new item parameters. The same constants were applied to the 1992 sample proficiency distribution. The transformation that was applied is as follows: $\theta = A\theta^{*} + B$ where θ* is the provisional scale from item calibration and θ is the reported 0 to 500 scale. Table 9-2 presents the transformation constants (that is, the standard deviations and means) for the distributions of the three scales. These constants apply both to the 1992 data, and to the 1985 data when the new item parameters are used.

Table 9-2. Transformation constants (standard deviations and means) by literacy scale, 1992 and 1985 (using new item parameters)

Literacy scale	A (standard deviations)	B (means)
Prose	51.67	269.16
Document	52.46	237.50
Quantitative	54.41	276.87

9.3.3 Evaluation of Differential Group Performance

Performance differences across subpopulations were examined by constructing empirical characteristic curves of tests rather than of items for major subpopulations defined by variables such as gender and ethnicity.

Yamamoto and Muraki (1991) have found that sets of estimated item parameters, each estimated on separate calibration samples with different racial/ethnic compositions, differed significantly even after an appropriate linear transformation was applied to account for the scale indeterminacy. This suggests differential item functioning (DIF) by racial/ethnic subpopulations. The National Adult Literacy Survey assessment as a whole functioned equivalently, however, suggesting that the effects of a different set of item parameters on the estimated proficiency of subpopulations may be negligible. In fact, after a linear scale transformation to account for the scale indeterminacy was applied to the real data, the estimates of subgroup proficiency distributions using a different set of item parameters were virtually identical. Since the main goal was to prevent systematic bias against any particular subpopulation, it was more appropriate to evaluate differential group performance at the test level than at the item level. Therefore, empirical test characteristic curves were constructed for the various sex, racial/ethnic, and age groups. These are shown in Exhibits 9-2p, 9-2d, and 9-2q, one for each scale.

The plots illustrate the average empirical proportion correct for the tasks in each literacy scale for each sex, racial/ethnic, and age group. Each point on the scale was estimated in two steps. First, the empirical proportion correct for every task was calculated for each sample for those whose proficiency values were in the selected 20-point range for at least one of 10 plausible values; second, the percents correct were then averaged for all tasks in the scale. This procedure was repeated for each subpopulation of interest. While the plot for document literacy scale by age groups (Exhibit 9-2d), and several others show deviations in the test characteristic curves within either the very low (below 200) and very high (above 360) parts of the proficiency ranges, the number of individuals performing in these ranges is very small, and therefore stable estimates cannot be made. Thus, when comparing test characteristic curves, one should concentrate on the part of the proficiency range where most of the population scores.

If the test characteristic curves deviated systematically within a subpopulation of interest, this could be viewed as evidence that the test is functioning differentially (is biased) for that group. The subpopulation curves were quite similar, however. Thus, it is safe to conclude that viewing the test as a whole, differential functioning was not observed across sex or racial/ethnic or age subpopulations in the National Adult Literacy Survey.

Exhibit 9-2p. Prose literacy test characteristic curves, by gender, race/ethnicity, and age: 1992

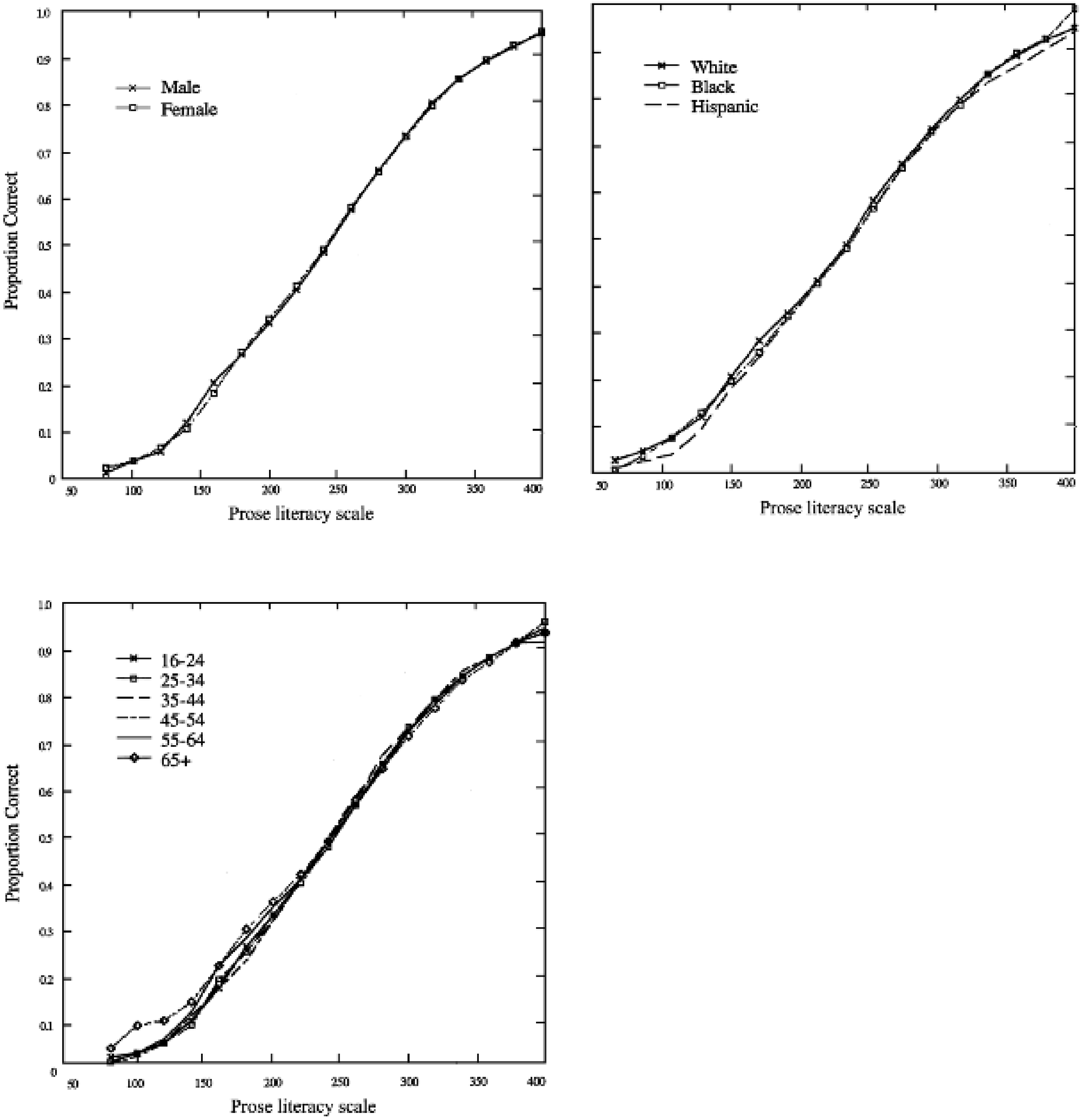


Exhibit 9-2d. Document literacy test characteristic curves, by gender, race/ethnicity, and age: 1992

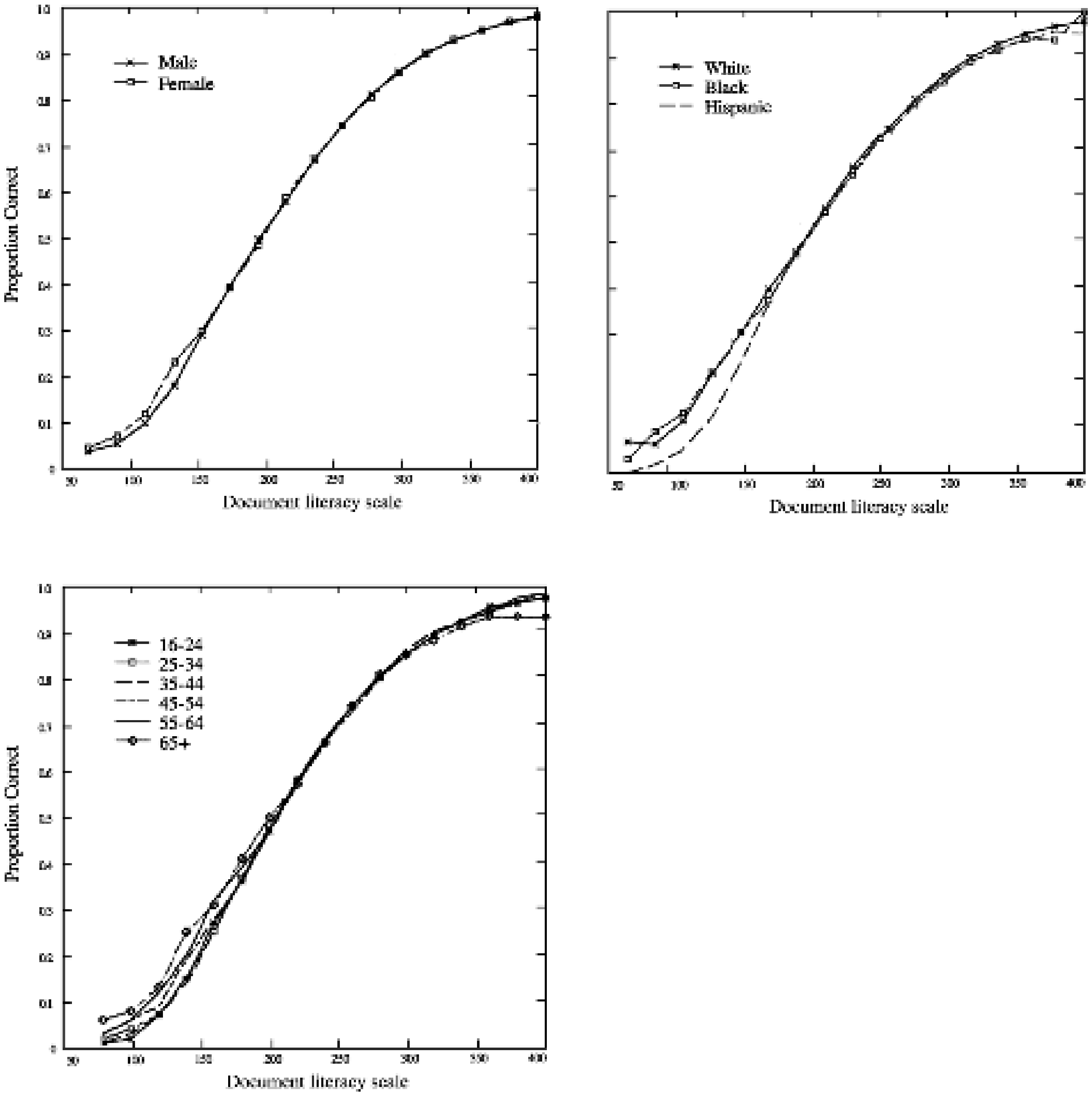
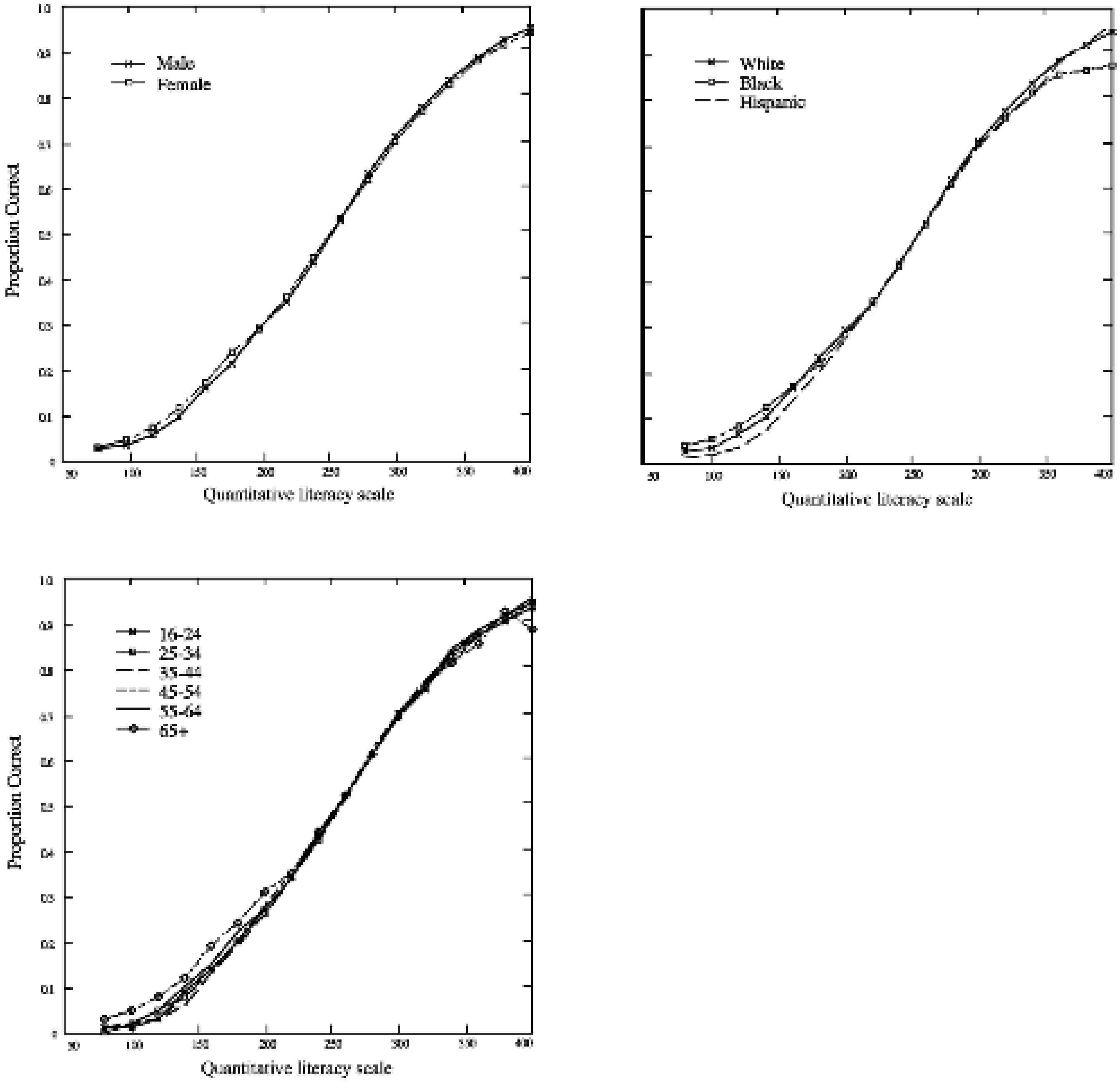


Exhibit 9-2q. Quantitative literacy test characteristic curves, by gender, race/ethnicity, and age: 1992



9.4 STATISTICAL TESTS

9.4.1 Analysis of Plausible Values

Plausible values methodology was used in this survey to increase the accuracy of the proficiency distribution estimates for various subpopulations and for the adult population as a whole. This method correctly retains the uncertainty associated with proficiency estimates for individual respondents by using multiple imputed proficiency values rather than assuming that this type of uncertainty is zero—a more common practice. Retaining this component of uncertainty requires that additional analysis procedures be used to estimate respondents’ proficiencies.

If the true θ values were observed for all sampled respondents, the statistic $\frac{t - T}{\sqrt{U}}$ would follow a t-distribution with d degrees of freedom. Since the true θ values are unknown, only incomplete data are available. The corresponding incomplete-data statistic $\frac{t^* - T}{\sqrt{\mathrm{Var}\,(t^*)}}$ is approximately t-distributed, with degrees of freedom given by

$$v = \frac{1}{\frac{f_M^2}{M - 1} + \frac{(1 - f_M)^2}{d}} \tag{9}$$

where f_M = the proportion of total variance due to not observing q values:

M = sets of plausible values

d = degrees of freedom associated with $\frac{t - T}{\sqrt{U}}$

$$f_M = \frac{\left(1 + \frac{1}{M}\right) B_M}{V_M} \tag{10}$$

where B_M = variance among the M estimates.

When B_M is small relative to U^* (average sampling variance over the M sets of plausible values), the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This was the case for the National Assessment of Educational Progress surveys. If, in addition, d is large, the normal approximation can be used instead of the t-distribution.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_M and U^* is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)V^{-1}(T-t^*)'$ is approximately F distributed with degrees of

$$f_M = \frac{(I + \ M^{-1})\,\mathrm{Trace}(B_M V_M^{-1})}{k} \qquad (11)$$

freedom equal to k and v, with v defined as above but with a matrix generalization of f_M

A chi-square distribution with k degrees of freedom can be used in place of f_M for the same reason that the normal distribution can approximate the t distribution.

Statistics t*, the estimates of ability and background variables, are consistent estimates of the corresponding population values T, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987). To avoid such biases, the National Adult Literacy Survey analysis included nearly all background variables, coded as dummy variables. To capture most of the variances in the background questions with a limited number of variables, principal components were used. Because each subpopulation can have unique relationships among the background variables, one set of principal components is not sufficient for all samples included in the National Adult Literacy Survey (i.e., the older adult, prison, and household samples). Each set of principal components was selected to include 99 percent of the variance in the background variables. Mislevy (1990) shows that this puts an upper bound of 1 percent on the average bias for all analyses involving the original conditioning variables.

9.4.2 Partitioning the Estimation Error Variance: A Numerical Example

This section offers an example of the use of multiple plausible values in the National Adult Literacy Survey analysis to partition the error variance. Table 9-3 presents data for three subgroups of respondents with differing educational attainments: those whose highest level of education was a GED, a high school diploma, and a four-year college degree. As noted earlier, five plausible values were calculated for each respondent for each scale. Each column presents the means of these five values.

Table 9-3. Mean plausible values by level of education for the prose scale

Level of Education	Sample N	Five imputed values					Mean	Var	JK_1 var	Standard error
		1	2	3	4	5				
GED	1062	269.3	268.1	267.9	268.2	267.7	268.2	0.483	2.888	1.84
High school	6107	270.2	270.4	270.3	270.5	270.2	270.3	0.180	1.050	1.11
4-year college	2534	321.2	321.7	322.4	322.8	320.4	321.7	1.027	1.408	1.56

Variance in the mean plausible values is similar but not identical for the three subgroups. As noted previously, variance reflects a component of error attributable to the measurement instrument's lack of precision and a component of error attributable to sample size. Variance can be reduced by either increasing the precision of the measurement instrument (for example, expanding the number of items) or increasing the size of the sample. The jackknife method was used to estimate error variance due to sampling using the first set of imputed values. This component of variance is expected to be consistent across the imputed values, and the size is influenced by the homogeneity of proficiencies among respondents in a subgroup but not by the sample size or by the precision of the survey instruments. Error variance due to sampling is smaller when the subgroup consists of respondents with similar proficiencies.

Despite a relatively large sample size, the mean for respondents with four-year college degrees has a larger error variance than those for other education groups. In fact, it is twice as large as the variance for respondents whose highest level of education is a GED. The higher variance for this best educated group is due to the characteristics of the assessment, which encompassed the entire adult population (age 16 and older) in this country and measured a wide range of skills. The precision of the assessment is optimal at the middle of the proficiency range, since that is where most of the population is expected to perform. Since the majority of the respondents with four-year college degrees scored above this range, variance due to lack of precision in measurement is quite high. Therefore, increasing the sample size would not do much to reduce the variance component for this group. On the other hand, the error variance due to sampling is twice as large for the smaller GED group as for the larger four-year college degree group.

The last column presents the standard error of the subpopulation mean, which is equal to the square root of the sum of the two components of error variance. The differences among the means can be compared using these standard errors. In doing so, it is first necessary to decide how many comparisons are being made. For this example, one might be interested in making three comparisons: GED vs. high school, high school vs. four-year college degree, and GED vs. four-year college degree. Following the Bonferroni method of multiple comparisons, any comparison among these three with a standardized difference greater than 2.39—$(\text{mean}_1 - \text{mean}_2)/\text{sqrt}(\text{se}_1^2 + \text{se}_2^2)$, $(z_p = 0.025/3)$—can be considered statistically significant. The difference in means between GED recipients and high-school graduates is not statistically significant at the .05 level, but the differences between these two groups and respondents with four-year degrees are significant.

9.4.3 Minimum Sample Sizes for Reporting Subgroup Results

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

9.4.4 Estimates of Standard Errors with Large Mean Squared Errors

Standard errors of mean proficiencies, percentages, and percentiles play an important role in interpreting subpopulation results and comparing the performances of two or more subpopulations. The jackknife standard errors reported for the National Adult Literacy Survey are statistics whose quality depends on certain features of the samples from which the estimates are obtained. In certain cases—primarily when the standard error is based on a small number of respondents—the mean squared error associated with the estimated standard errors may be quite large. In the survey reports, estimated standard errors that are subject to large mean squared errors are followed by the symbol "!", indicating that the coefficient of variation (CV) is greater than 0.2. This CV is estimated by:

$$\mathrm{CV}(\hat{\mathrm{N}}) = \frac{\mathrm{SE}(\hat{\mathrm{N}})}{\hat{\mathrm{N}}} \tag{12}$$

where $\hat{\mathrm{N}}$ is a point estimate of N and SE($\hat{\mathrm{N}}$) is the jackknife standard error of $\hat{\mathrm{N}}$.

Experience with other large-scale assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means, and percentages based on samples of this size, may be quite large. Therefore, these standard errors, and any confidence intervals or significance tests involving them, should be interpreted with caution. Johnson and Rust (1992) discuss this issue in detail.