Journal of Educational Statistics
Summer 1992, Vol. 17, No. 2, pp. 131–154

Chapter 3

Scaling Procedures in NAEP

Robert J. Mislevy, Eugene G. Johnson, and **Eiji Muraki**
Educational Testing Service

Key words: *item response theory, multiple imputation, NAEP*

Scale-score reporting is a recent innovation in the National Assessment of Educational Progress (NAEP). With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale even when different students have been administered different exercises. This article presents an overview of the scaling methodologies employed in the analyses of NAEP surveys beginning with 1984. The first section discusses the perspective on scaling from which the procedures were conceived and applied. The plausible values methodology developed for use in NAEP scale-score analyses is then described, in the contexts of item response theory and average response method scaling. The concluding section lists milestones in the evolution of the plausible values approach in NAEP and directions for further improvement.

NAEP reports were originally envisaged some 20 years ago as simple lists of percents correct to individual survey items, in the population as a whole and in subpopulations of particular interest. It soon became apparent that major features of the detailed results from hundreds of items could not be effectively communicated without some kind of summarization. Averaging percents correct from individual items aggregates results but limits comparisons to groups of items that are common over the time points or student subpopulations that are to be compared.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in

The contributions of Nancy Allen, Albert E. Beaton, David S. Freund, Bert F. Green, Huynh Huynh, Bruce A. Kaplan, Jo-Ling Liang, Bengt O. Muthén, Jennifer Nelson, Kathleen M. Sheehan, Neal Thomas, Minhwei Wang, Kentaro Yamamoto, and Rebecca Zwick to this article are gratefully acknowledged. The work on which this article is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. When combined through the appropriate scaling formulas, these variables capture the dominant features of the data. Using the scale, it becomes possible to talk about distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables. Item response theory (IRT; see Hambleton, 1989, for an overview) and a newly developed procedure called *the average response method* (ARM), both of which are reviewed below, are the two scaling procedures that have been employed in NAEP reporting to date.

Of course, any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. In a very real sense, every single item in a NAEP survey is of interest in its own right and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose. The procedure that is optimal for one purpose may be poorly suited for another. The relatively high levels of aggregation found in NAEP reports such as *The Reading Report Card: Progress Toward Excellence in Our Schools* (NAEP, 1985), for example, are suited to general discussions of trends and policy implications. These reports average over, and are therefore not keyed to, patterns of performance at the level of specific skills; they do not reveal popular student misconceptions or erroneous rules that might be of interest to classroom teachers in a subject area. By no means do the scale-score methods one employs as a reporting vehicle exhaust the potential of NAEP data. NAEP secondary-use data files, which contain the original responses of all surveyed students, were created expressly to encourage secondary researchers to carry out alternative analyses from different perspectives.

The reporting scale of a NAEP survey, then, simply summarizes performance on a collection of educational tasks in much the same way that the Consumer Price Index (CPI) summarizes the cost of a market basket of products. The two indexes exhibit some of the same useful features and limitations. Just as the CPI composite represents average American spending patterns, the items in a NAEP survey were specified by independent consensual process to tap a market basket of skills. Just as changes in the CPI reflect at a glance changes in the cost of goods in general, changes in NAEP scale-score distributions reflect changes in proficiency as averaged over the items in the pool. Understanding just how and why the CPI changes requires deeper analysis into specific components of the market basket; when the CPI goes up, some of the components will have gone up by greater rates than others, while others may have dropped. The NAEP scale depends similarly on the balance of items of varying types and topics in the survey and reflects only an average over the varying patterns among them.

NAEP first attempts to carry out scaling in subject areas in which similar patterns can be expected over items—for example, within six more narrowly defined topics within mathematics such as algebra, measurement, and numbers and operations. Carrying out scaling in separate subareas captures trends and comparisons that may differ across subareas because of different curricular emphases over time or across schools. Subscale results are supplemented by a subject-area average in surveys such as those of mathematics and science in 1986. This is comparable to calculating price changes in separate market baskets for food, transportation, energy, and so on and reporting these along with the overall average. Within each scaling area, meaningful departures from general trends are additionally highlighted by supplementing scale-score distributional results with more detailed breakdowns, in terms of percents correct for groups of related items, and explicating countertrends or comparisons that can be identified with one or a few items or with particular subpopulations of students. This is analogous to reporting that the Consumer Price Index jumped 5% but noting that the increase was mainly due to a change in OPEC oil prices.

NAEP Scaling Methodology

This section reviews the scaling models employed in the analyses of NAEP data beginning in 1984 and the plausible values methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy and Sheehan (1987) and Beaton and Johnson (1987, 1990) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach.

The Scaling Models

Two types of scaling models have been used by NAEP in recent assessments. The three-parameter logistic (3PL) model from item response theory (IRT; see Lord, 1980) was used for the subject areas of reading, civics, U.S. history, geography, mathematics, and science. The average response method (ARM; Beaton & Johnson, 1987, 1990), an extension of multiple regression developed by NAEP for the 1984 assessment, was used for the subject area of writing and for summarizing background information and attitude responses. The 3PL and the ARM are both latent variable models, quantifying respondents' tendencies to provide responses in a given direction (e.g., correct answers to items in a subject area, positive responses on attitude questions, higher rather than lower ratings in written essays) as a function of a parameter that is not directly observed.

The three-parameter logistic IRT model. The fundamental equation of the 3PL model is the probability that a person whose proficiency is characterized by the unobservable variable θ will respond correctly to item j:

$$P(x_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(\theta - b_j)]\}$$
$$\equiv P_j(\theta), \tag{1}$$

where

x_j is the response to item j, 1 if correct and 0 if not;
a_j, where $a_j > 0$, is the slope parameter of item j, characterizing its sensitivity to proficiency;
b_j is the threshold parameter of item j, characterizing its difficulty; and
c_j, where $0 \leq c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of a correct response from students of very low proficiency. In NAEP analyses, c parameters are estimated for multiple-choice items but fixed at zero for open-ended items.

For the purposes of reporting item parameter estimates and other intermediate estimates, the linear indeterminacy apparent in (1) may be resolved by an arbitrary choice of the origin and unit size in a given scale. For the 1984 reading analyses, for example, calculations were carried out using a scale that standardized the combined grade 4/age 9, grade 8/age 13, and grade 11/age 17 samples. To facilitate interpretation, final estimates were reported on a 0–500 scale related virtually linearly to the θ scale between -4 and $+4$. Scaling conventions are described in detail in corresponding subject area chapters in NAEP technical reports (Beaton, 1987, 1988; Johnson & Allen, 1992; Johnson & Zwick, 1990).

In NAEP, estimates of item parameters were obtained with a modified version of Mislevy and Bock's (1982) BILOG computer program and then treated as known in subsequent calculations. Once items have been calibrated, a likelihood function for θ is induced by a vector of responses to any subset of calibrated items, thus allowing inference on the θ scale from matrix samples.

Under the usual IRT assumption of conditional, or local, independence, the probability of a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ of responses to n items is simply the product of terms based on (1):

$$P(\boldsymbol{x} \mid \theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_j^n [P_j(\theta)]^{x_j} [1 - P_j(\theta)]^{1-x_j}. \tag{2}$$

It is typically also assumed that response probabilities are conditionally independent of background variables—say, $\boldsymbol{y}$, given θ, or

$$P(\boldsymbol{x} \mid \theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{y}) = P(\boldsymbol{x} \mid \theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}).$$

Conditional independence is a mathematical assumption, not a fact of nature. Even though the IRT models are employed in NAEP only to summarize average performance, several checks are made to detect serious violations of conditional dependence, and, when warranted, remedial efforts are made to mitigate the effects of these violations on inferences. Checks on the plausibility of the scaling assumptions include the following:

(i) Item operating characteristics among distinct gender and ethnicity groups are compared (i.e., differential item functioning [DIF] analyses). Some degree of relative differences is to be expected, of course, and modestly varying profiles among groups will exist beyond the differences conveyed by their differing θ distributions. The intent is to detect and eliminate items that operate differentially for identifiable reasons unrelated to the skills intended to be measured.

(ii) When a subscale extends over age groups, evidence is sought of different operating characteristics over ages. When such effects are found, an item in question is represented by different item parameters in different age groups. For such an item, the probability of a correct response, given θ, depends on the age group in question—a departure from conditional independence incorporated into the model in the interest of fidelity to the data. This is analogous to calibrating items separately in different age groups and linking the resulting scales via those items whose response curves in the separate ages can be rectified by a single linear transformation.

(iii) When a scale extends over time, evidence is similarly sought as to whether an item's relative operating characteristics have changed over time, over and above differences that can be accounted for by changes in the overall θ distribution. Studies of NAEP reading data reported in Beaton and Zwick (1990) suggest these effects may be relatively small in adjacent assessments when assessment forms are held constant, but they too can be taken into account by modeling atypical items as different items at different time points, as discussed above.

Another type of check, item-level factor analysis, has diminished in importance as our perspective of the role of IRT in NAEP has evolved. The assumption that performance in a scaling area is driven by a single unidimensional variable is unarguably incorrect. Our use of the model is not theoretic but data analytic; interpretation of results is not trait referenced but domain referenced. Scaling areas are determined a priori by considerations of content and politics: NAEP subscales are collections of items for which overall performance is deemed to be of interest. The IRT summary is expected not to capture all meaningful variation in item response data but to reflect distributions of overall proficiency—to summarize the main patterns in item percents correct in the populations and subpopulations of interest. Attention turns to avoiding or mitigating violations that distort the most important inferences that are to be drawn.

Estimated proficiency distributions and item parameters, for example, correctly capture groups' overall performances as averaged over items but, while doing so, may overestimate performance on certain items and underestimate performance on others. Because many reports emphasize average results, the composition of the item pool in a scale is more important than it would be if the IRT model were true, because the mix of items influences the nature of the average proficiency that scale scores convey. Item mixes are identical for subpopulations within a given assessment, so that overall proficiency has essentially the same meaning for them. Item mixes are not necessarily the same for different age groups or different time points, so that the checks discussed above are required to maintain, as well as possible, meaning across these linkages.

The assumption of local independence additionally implies that item response probabilities do not depend on such factors as the position of the item in the booklet, the content of items around an item of interest, or test-administration timing conditions. These effects are certainly present in any application, however. The practical question is whether the IRT probabilities obtained via (2) are close enough to be robust with respect to the context in which the data are to be collected and the inferences that are to be drawn. In the 1986 NAEP reading anomaly, for example, changes in item context and speededness conditions led to unacceptably large random error components for measuring small changes in population means over time (Beaton & Zwick, 1990). These can be avoided by presenting items used to measure change in identical test forms with identical timings and administration conditions. Thus, we do not maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations, and we acknowledge that the parameter estimates are context-bound.

In all NAEP IRT analyses, missing responses at the end of each block were considered not reached and were treated as if they had not been presented to the respondent. This common IRT practice can introduce slight biases into item parameter estimation to the degree that not-reached items are present, and speed is correlated with proficiency (Mislevy & Wu, 1988). Missing responses before the last observed response in a block were considered intentional omissions and treated as fractionally correct at the value of the reciprocal of the number of response alternatives. Handling omissions in this manner provides consistent limited-information likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly (Mislevy & Wu, 1988).

The 3PL pertains to dichotomous, or right/wrong, test items. An increasing number of NAEP tasks, however, yields data in the form of ratings—for example, ratings of 0–4 for quality of the performance. Trial analyses with IRT models for such data are in progress.[1] Through the 1990 assessment, relatively few of these kinds of items appeared in subject areas addressed mainly through multiple choice, such as reading, mathematics, and science. Ratings for these items were collapsed into dichotomies for IRT analyses. In writing, all items provided rating data. Writing analyses were carried out using the method described below.

The average response method model. In the 1984 and 1988 NAEP surveys, the average response method (ARM) was used to carry out analyses of writing data. The ARM writing composite variable is defined as an average rating (on the 0–4 rating scale for responses to essay prompts) over the set of essay exercises. Under the NAEP balanced incomplete block-spiraled (BIB-spiraled) sampling design, a given student is administered only between one and four of the prompts. ARM analyses permit summary reports to be described in terms of students' projected average performance on the entire set of prompts. The basic equation is an average of item responses:

$$\theta = \boldsymbol{a}'\boldsymbol{x} \,. \tag{3}$$

Here, $\boldsymbol{a}$ is a vector of constants, specified to provide a meaningful summary of performance. Weights of $1/p$ for a p item test, for example, yield simply an average score. Note that, if a respondent responded to all items, an ARM score would be directly calculable via (3) without error. Computational details will be provided in a following section.

An Overview of Plausible Values Methodology

IRT was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 50 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can be approximated by using individuals' $\hat{\theta}$ values as if they were θ values. This approach breaks down in the large-scale assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The uncertainty associated with individual θs is too large to ignore, and, paradoxically, point estimates of θ that are optimal for each individual examinee can have a distribution that is seriously biased as an estimate of the θ distribution (Little & Rubin, 1983). We note in passing that the same paradox arises in the estimation of change (Cronbach & Furby, 1970) and of factor scores (Tucker, 1971).

Consistent estimates of population characteristics can be obtained nevertheless using marginal estimation techniques that bypass the step of estimating scores for individuals—by directly maximizing likelihood functions that involve only the population parameters and the raw data (e.g., see Dempster, Laird, & Rubin, 1977; Mislevy, 1984, 1985). This is done in NAEP; no scores are estimated for individual respondents. NAEP does, however, provide plausible values to represent the distributions. A file of plausible values of latent variables is constructed to yield, when analyzed as if it contained the true values of those latent variables, the population characteristics obtained in the marginal analysis. A detailed development of plausible values methodology is given in Mislevy (1991). The following paragraphs give an overview of the approach.

Let $\boldsymbol{Y}$ represent the responses of all sampled examinees to background and attitude questions, with design variables such as school membership. If IRT or ARM θ values were available for all sampled examinees, it would be possible to compute a statistic $t(\boldsymbol{\theta}, \boldsymbol{Y})$—such as a subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T. Another function $U(\boldsymbol{\theta}, \boldsymbol{Y})$—for example, a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the 3PL model and the ARM are latent variable models, however, θ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by thinking of θ as missing data and approximating $t(\boldsymbol{\theta}, \boldsymbol{Y})$ by its expectation given $(\boldsymbol{X}, \boldsymbol{Y})$, the data that actually were observed:

$$\begin{aligned} t^*(\boldsymbol{X}, \boldsymbol{Y}) &= E[t(\boldsymbol{\theta}, \boldsymbol{Y}) \mid \boldsymbol{X}, \boldsymbol{Y}] \\ &= \int t(\boldsymbol{\theta}, \boldsymbol{Y})\, p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})\, d\boldsymbol{\theta}. \end{aligned} \tag{4}$$

In special cases, it is possible to calculate the integral equation in (4) directly, thus obtaining an estimate of a population characteristic without ever obtaining a score estimate for a single individual. Closed-form solutions are not forthcoming with IRT measurement models, however, so that alternative methods must be sought to evaluate (4). Stochastic, or Monte Carlo, integration uses random draws from the conditional distributions $p(\theta \mid \boldsymbol{x}_i, \boldsymbol{y}_i)$ of each sampled student i. This approximation yields intermediate calculations with connections to missing-data analysis: The random draws can be viewed as *imputations,* in the terminology of the sampling literature. They are the NAEP plausible values. The value of θ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from the conditional distribution for θ given the respondent's responses to cognitive items $(\boldsymbol{x}_i)$ and background items $(\boldsymbol{y}_i)$. Rubin (1987) proposes that this process be carried out several times—multiple imputations—so that the uncertainty associated with the fact that θs are not observed can be quantified. The average of the results of, say, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (4). The variance among them reflects uncertainty due to not observing θ and must be added to an estimate of $U(\boldsymbol{\theta}, Y)$, which reflects uncertainty due to testing only a sample of students.

Note from (4) that plausible values are drawn from distributions that already convey implicitly the characteristics of the population through the factor $p(\boldsymbol{\theta} \mid X, Y)$. This is obtained through the marginal analyses described below, which do not require scores for individual respondents. The plausible values thus only reflect the population characteristics with which they were constructed, as estimated in the marginal analyses. The steps in this process reverse those employed when precise scores are available for each respondent, because in those more familiar situations one can estimate population characteristics from individuals' scores. Because plausible values are offered only as intermediate calculations to compute estimates of population characteristics via (4), and not for inferences or decisions about individual students, it cannot be emphasized too strongly that plausible values are not test scores for individuals in the usual sense.

Computing Plausible Values in IRT-Based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\theta \mid \boldsymbol{x}_i, \boldsymbol{y}_i)$. This subsection describes how, in IRT-based scales, these conditional distributions are characterized and how the draws are taken. Using first Bayes' theorem, then the IRT assumption of conditional independence (i.e., $P(\boldsymbol{x}_i \mid \theta, \boldsymbol{y}_i) = P(\boldsymbol{x}_i \mid \theta)$),

$$\begin{aligned} p(\theta \mid \boldsymbol{x}_i, \boldsymbol{y}_i) &\propto P(\boldsymbol{x}_i \mid \theta, \boldsymbol{y}_i)\, p(\theta \mid \boldsymbol{y}_i) \\ &= P(\boldsymbol{x}_i \mid \theta)\, p(\theta \mid \boldsymbol{y}_i), \end{aligned} \tag{5}$$

where $P(\boldsymbol{x}_i \mid \theta)$ is the likelihood function for θ induced by observing $\boldsymbol{x}_i$ (treating item parameter estimates as known true values) and $p(\theta \mid \boldsymbol{y}_i)$ is the distribution of θ given the observed value $\boldsymbol{y}_i$ of background responses.

Equations 4 and 5 can also be employed with vector-valued $\boldsymbol{\theta}$. In cases such as in the 1986 NAEP mathematics subscales, where each item is assumed to depend on only one subscale proficiency dimension, the multivariate likelihood $P(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ is the product over subscales of the independent likelihoods induced by responses to items within each subscale, and $p(\boldsymbol{\theta} \mid \boldsymbol{y}_i)$ is the multivariate—and generally nonindependent—joint density of proficiencies for the subscales, conditional on background variables $\boldsymbol{y}$.

In the analyses of NAEP data from 1986 onward, a normal (Gaussian) form was assumed for $p(\boldsymbol{\theta} \mid \boldsymbol{y}_i)$, with a common dispersion and with a mean given by a linear model for selected main effects and two-way interactions of the complete vector of background variables. The included background variables and interactions will be referred to as *conditioning variables* and will be denoted y^c. The following model was fit in each subject area:

$$\boldsymbol{\theta} = \boldsymbol{\Gamma}' y^c + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{\epsilon}$ is normally distributed with mean $\mathbf{0}$ and dispersion $\boldsymbol{\Sigma}$. $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are the parameters to be estimated. In subject areas with only one scale, such as reading, $\boldsymbol{\Gamma}$ is a vector, and $\boldsymbol{\Sigma}$ is a scalar. In subject areas comprising subscales, $\boldsymbol{\Gamma}$ is a matrix, and $\boldsymbol{\Sigma}$ is a covariance matrix. As in regression analysis, $\boldsymbol{\Gamma}$ is a vector or *matrix of effects,* and $\boldsymbol{\Sigma}$ is the scalar or *matrix variance of residuals.* Also, as in regression, the interpretation of the effects depends on how the design vectors in y^c are coded—as contrasts or for linear effects, as examples. For instance, four regions of the country were included in the analysis. Three region effects were estimated, corresponding to simple contrasts of the Northeast region with each of the others. Like item parameter estimates, the estimates of the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ of conditional distributions were treated as known true values in subsequent steps of the analyses.

Maximum likelihood estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ were obtained with Sheehan's (1985) M-GROUP computer program, using a variant of the EM solution described in Mislevy (1985). The difference lies in the numerical approximation employed for examinees' likelihood functions, now described. Note from (5) that $p(\boldsymbol{\theta} \mid \boldsymbol{x}_i, y_i)$ is proportional to the product of two terms, the likelihood $P(\boldsymbol{x}_i \mid \boldsymbol{\theta})$, and the conditional distribution $p(\boldsymbol{\theta} \mid y_i)$. The conditional distribution has been assumed multivariate normal, with mean $\boldsymbol{\mu}_i^c = \boldsymbol{\Gamma}' y_i^c$ and covariance matrix $\boldsymbol{\Sigma}$; if the likelihood is approximated by another normal distribution, with mean $\boldsymbol{\mu}_i^L$ and covariance matrix $\boldsymbol{\Sigma}_i^L$, then the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_i, y_i)$ is also multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_i^P = (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}_i^L)^{-1})^{-1} \tag{7}$$

and mean vector

$$\tilde{\boldsymbol{\theta}}_i = (\boldsymbol{\theta}_i^c \boldsymbol{\Sigma}^{-1} + \boldsymbol{\theta}_i^L (\boldsymbol{\Sigma}_i^L)^{-1}) \boldsymbol{\Sigma}_i^P. \tag{8}$$

(See Box & Tiao, 1973, appendix A1.1, for a derivation in the univariate case.) The likelihood induced by a respondent's answers to the items in a given scale is thus approximated by a normal distribution—univariate normal distributions in subject areas with only one scale, multiple independent normal distributions in an area comprising multiple scales.

This normalized-likelihood/normal posterior approximation was employed in both the estimation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ and in the generation of plausible values. A plausible value was drawn at random from this normal distribution—again univariate normal for subject areas with only a single scale, multivariate normal for those with multiple subscales. For subject areas with multiple subscales, weighted-average composites over subscales were also calculated after appropriate rescaling (see Beaton, 1988, for details and definitions of composites).

Computing Plausible Values in ARM Scales

Recall that the average response method begins with a defined composite of scores for a set of p exercises where each exercise is rated on a multiple-point scale; that is, $\theta = \boldsymbol{a}'\boldsymbol{x}$, where $\boldsymbol{x}$ is the vector of the subject's scores on the p exercises in the composite and $\boldsymbol{a}$ is a vector of p predetermined constants. Under the NAEP BIB-spiraled item sampling design, each student is presented only a subset of the exercises, so that each student's composite value is not directly observed. As shown below, however, it is possible nevertheless to estimate consistently the covariance among each pair of exercises. Estimating the mean vector and the covariance matrix among exercises ($\boldsymbol{x}$) and conditioning variables ($\boldsymbol{y}$) constitutes the marginal analysis phase of the ARM. Like IRT plausible values, ARM plausible values are drawn at random from respondents' predictive distributions $p\,(\theta\,|\,\boldsymbol{x},\boldsymbol{y})$ and serve not as estimates of proficiencies of individual students but as intermediate computations for estimating population characteristics.

Let $\boldsymbol{x}_i$ represent the (row) vector of responses of the ith student to the exercises in the ARM composite that were presented to that student, and let $\boldsymbol{y}_i$ be the (row) vector of values of that student's conditioning variables. For convenience, we treat the ARM composite as a continuous variable; it is in fact discrete, but it can take on a large number of closely spaced values. We approximate the predictive distribution of the composite score of student i by

$$p\,(\theta\,|\,\boldsymbol{x}_i,\boldsymbol{y}_i) = n\,(\boldsymbol{x}_i\,\hat{\boldsymbol{\beta}} + \boldsymbol{y}_i\,\boldsymbol{\Gamma}, \sigma_\epsilon^2),$$

approximating the distribution of a residual term ϵ by a normal distribution. A plausible value from the conditional distribution of θ, given the observed data $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ for student i, is then constructed as

$$\tilde{\theta}_{ik} = \boldsymbol{x}_i\,\hat{\boldsymbol{\beta}} + \boldsymbol{y}_i\,\hat{\boldsymbol{\Gamma}} + \boldsymbol{x}_i\,\boldsymbol{\alpha}_k + \boldsymbol{y}_i\,\boldsymbol{\gamma}_k + \epsilon_{ik}, \qquad (9)$$

where

$\tilde{\theta}_{ik}$ is student i's kth plausible value of the ARM composite;

$\hat{\boldsymbol{\beta}}$ is a (column) vector giving the change in the composite for unit change in the scores on each of the exercises in $\boldsymbol{x}_i$;

$\hat{\boldsymbol{\Gamma}}$ is a (column) vector of effects for the conditioning variables $\boldsymbol{y}_i$;

$\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$ are random draws from a multivariate normal distribution with mean vector $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Gamma}}$ ($\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$ reflect the uncertainty due to using sample estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Gamma}}$ in the regression equation); and

ϵ_{ik} is an estimated residual drawn from a normal distribution with mean 0 and variance σ_ϵ^2, the variance of the predictive distribution of θ given the observed values of $\boldsymbol{x}_i$ and y_i.

All parameters in (9) can be estimated consistently by least squares. It suffices to obtain estimates of the elements of the population sum of squares and cross-products matrix of the conditioning variables and the exercises:

$$\boldsymbol{C} = \text{an estimate of } \boldsymbol{V}'\boldsymbol{V} = \begin{bmatrix} \boldsymbol{Y}'\boldsymbol{Y} & \boldsymbol{Y}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Y} & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix};$$

$\boldsymbol{Y}$ is a $N \times q$ matrix containing the values of the q conditioning variables for each of the N students in the population; $\boldsymbol{X}$ is a $N \times p$ matrix containing the scores of the N students in the population on the p exercises, and $\boldsymbol{V} = [\boldsymbol{Y}\,\boldsymbol{X}]$. If $\boldsymbol{Y}$ and $\boldsymbol{X}$ were known for all students in the population, $\boldsymbol{C}$ would be trivially equal to $\boldsymbol{V}'\boldsymbol{V}$. However, because only a sample of the students in the population was assessed and because each sampled student was only presented a few of the exercises, many of the elements of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are unknown. We will return shortly to the procedures used to determine an estimate $\boldsymbol{C}$ of $\boldsymbol{V}'\boldsymbol{V}$.

Because the ARM composite is the mean of the individual exercises, the estimate $\boldsymbol{C}$ generates a complete set of sufficient statistics (the normal equations) for the standard least-squares prediction of an ARM composite value, given conditioning variable characteristics and responses to any subset of exercises. Define the N element column vector $\boldsymbol{T}$ by

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{a},$$

where the elements of $\boldsymbol{T}$—namely, θ_i for $i = 1, \ldots, N$—are the values of the composite for each student in the population. The exact value of θ_i will not be known unless the student i was administered all p of the exercises. The plausible values, $\tilde{\theta}_{ik}$, of (9) are determined by operations on the matrix $\boldsymbol{C}_\theta$, where $\boldsymbol{C}_\theta$ is the estimated population sum of squares and cross-product matrix of the conditioning variables, the exercises, and the composite. $\boldsymbol{C}_\theta$ is generated by transforming $\boldsymbol{C}$ as follows:

$$\boldsymbol{C}_\theta = \begin{bmatrix} \boldsymbol{I}_q & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p & \boldsymbol{a} \end{bmatrix}' \boldsymbol{C} \begin{bmatrix} \boldsymbol{I}_q & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p & \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}'\boldsymbol{Y} & \boldsymbol{Y}'\boldsymbol{X} & \boldsymbol{Y}'\boldsymbol{T} \\ \boldsymbol{X}'\boldsymbol{Y} & \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{T} \\ \boldsymbol{T}'\boldsymbol{Y} & \boldsymbol{T}'\boldsymbol{X} & \boldsymbol{T}'\boldsymbol{T} \end{bmatrix}.$$

$\boldsymbol{C}_\theta$ can be used to estimate appropriate values for student i as follows. Let $\boldsymbol{X}_1$ consist of the columns of $\boldsymbol{X}$ corresponding to the items presented to student i, and let $\boldsymbol{V}_1 = [\boldsymbol{Y}\boldsymbol{X}_1]$. The least-squares estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Gamma}}$ in (9) are

$$\begin{bmatrix} \hat{\boldsymbol{\Gamma}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}'\boldsymbol{Y} & \boldsymbol{Y}'\boldsymbol{X}_1 \\ \boldsymbol{X}_1'\boldsymbol{Y} & \boldsymbol{X}_1'\boldsymbol{X}_1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Y}'\boldsymbol{T} \\ \boldsymbol{X}_1'\boldsymbol{T} \end{bmatrix},$$

whence the standard least-squares point estimate of the composite score for student i is:

$$\hat{\theta}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}} + \boldsymbol{y}_i \hat{\boldsymbol{\Gamma}}.$$

This is the mean of the predictive distribution of θ for student i. As seen in (9), ARM plausible values are constructed by adding to this mean disturbance terms that convey remaining uncertainty of two types: (a) ϵ_{ik}, accounting for variability of potential scores of an individual about the conditional mean (of the distribution given $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$), and (b) $\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$, accounting for uncertainty due to using sample estimates of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\Gamma}}$ in the regression equation. ϵ_{ik} is a random draw from a $N(0, \sigma_\epsilon^2)$ distribution, where σ_ϵ^2 is the residual mean squared error for the regression defined by (10). The vector $(\boldsymbol{\gamma}_k, \boldsymbol{\alpha}_k)'$ is a draw from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\boldsymbol{\Sigma} = (V_1' V_1)^{-1} \sigma_\epsilon^2.$$

The values of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$ are held fixed for all students with the same pattern of missing data.

A further discussion of the generation of ARM plausible values, given an estimate $\boldsymbol{C}$ of $\boldsymbol{V}'\boldsymbol{V}$, appears in Beaton and Johnson (1990). We now consider the estimation of $\boldsymbol{V}'\boldsymbol{V}$.

As noted above, the basis for the estimation of a predictive distribution for any student is an estimate $\boldsymbol{C}$ of the full sums-of-squares-and-cross-products matrix $\boldsymbol{V}'\boldsymbol{V}$, from which all other necessary matrices and estimates are derived. The values of the conditioning variables are known for all students, and so the $\boldsymbol{Y}'\boldsymbol{Y}$ submatrix in $\boldsymbol{C}$ is directly obtained by taking the sum of squares and cross-products of the conditioning variables for each student, by weighting these by the student's sampling weight, and then by summing across all students of the given grade and assessment year.

The estimation of the remaining elements of the matrix $\boldsymbol{C}$ is most straightforward when consistent estimates of the population mean and variance of each exercise and the covariance between every pair of exercises are available. Such estimates are readily available when, as with the BIB-spiral design, every pair of exercises in the composite is presented to a representative sample of each population of interest. Specifically, if exercise j were presented to a representative sample of students of each population of interest, then the appropriately weighted sample mean, $\overline{X}_j$, and the weighted sample variance, S_j^2, based on the total sample of students, are unbiased estimates of the population mean and variance for the exercise. A consistent estimator of the sum-of-squared scores in the population for the exercise is

$$\widehat{\mathbf{X}_j' X_j} = W_{\text{TOT}}(S_j^2 + \overline{X}_j^2),$$

where W_{TOT} is the sum of weights for all students responding to any of the items going into the composite. A consistent estimate of the cross-product element $X_j' X_k$, when a representative sample of students were presented both exercises, is

$$\widehat{\mathbf{X}_j' X_k} = W_{\mathrm{TOT}}(S_j S_k r_{jk} + \overline{X}_j \overline{X}_k),$$

where S_j and $\overline{X}_j$ are the weighted sample standard deviation and mean based on the full sample of students responding to exercise j, S_k and $\overline{X}_k$ are the analogous statistics based on the full sample of students responding to exercise k, and r_{jk} is the appropriately weighted sample correlation coefficient based on the students responding to both exercises. The estimation of cross-products between conditioning variables and exercises, $\boldsymbol{Y}' \boldsymbol{X}_j$, is calculated analogously.

Although the estimates of each element of $\boldsymbol{V}' \boldsymbol{V}$ so obtained are consistent, the resulting matrix $\boldsymbol{C}$ need not be positive definite. The calculations described below that yield an individual's predictive distribution require inverting only selected submatrices and can be carried out nevertheless. A consistent positive definite estimate of $\boldsymbol{V}' \boldsymbol{V}$ can be calculated via maximum likelihood, using Dempster, Laird, and Rubin's (1977) EM algorithm. Empirical studies with the 1984 writing data showed that the predictive distributions obtained with ML estimates of $\boldsymbol{V}' \boldsymbol{V}$ were virtually indistinguishable from those obtained with $\boldsymbol{C}$.

By making additional assumptions, estimates of the elements of $\boldsymbol{C}$, and hence of the ARM composite, can be obtained when not all of the exercises or pairs of exercises are presented to each population of interest. Such a situation occurs when the performance of two populations is to be compared on the basis of a composite across a set of exercises, where not all exercises are presented to each population. For example, it may be desirable to compare the performance of grade eight students at one point in time with that of grade eight students assessed 2 years later using, as a measure of performance, a composite across p exercises where some of the exercises (linking exercises) were presented to both populations but where the remaining exercises (unique exercises) were only presented to one of the populations. The ARM can be extended to handle this situation by assuming that the conditional distribution of performance on the unique exercises, given the conditioning variables and the performance on the linking exercises, is the same for each of the populations and by making certain distributional assumptions about the correlations between pairs of exercises not presented together in any population. Johnson (1990) details this extension of the ARM.

Analyses

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source—namely, the sampling of respondents.

Item percents correct for NAEP cognitive exercises meet this requirement, but scale-score proficiency values do not. The IRT and ARM models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the area. The fact that θ values are not observed even for the respondents in the sample requires additional statistical machinery to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, we have adapted Rubin's (1987) multiple imputations procedures to the context of latent variable models to produce the plausible values on which many analyses of NAEP data are based. This section describes how plausible values are employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

Computational Procedures

Although we do not observe the θ value of respondent i, we do observe $\boldsymbol{x}_i$, the respondent's answers to the cognitive items he or she was administered in the area of interest, and $\boldsymbol{y}_i$, the respondent's values for demographic and background variables. Suppose that, if θ values had been observed, we had estimated a function T of the population values of θ and y using a sample statistic $t(\boldsymbol{\theta}, \boldsymbol{Y})$ [where $(\boldsymbol{\theta}, \boldsymbol{Y}) \equiv (\theta_1, y_1, \ldots, \theta_N, y_N)$] and we had estimated the variance in t around T due to sampling respondents by the function $U(\boldsymbol{\theta}, \boldsymbol{Y})$. Because observations consist of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ rather than $(\boldsymbol{\theta}_i, \boldsymbol{y}_i)$, we could approximate t by its expected value given $(\boldsymbol{X}, \boldsymbol{Y})$, or

$$t^*(\boldsymbol{X}, \boldsymbol{Y}) = E[t(\boldsymbol{\theta}, \boldsymbol{Y}) \mid \boldsymbol{X}, \boldsymbol{Y}] = \int t(\boldsymbol{\theta}, \boldsymbol{Y}) p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) \, d\boldsymbol{\theta}.$$

We now approximate t^* with random draws from the conditional distributions $p(\theta_i \mid \boldsymbol{x}_i, \boldsymbol{y}_i)$, as obtained for all respondents by the methods described above. The following steps describe how an estimate of a scalar statistic, $t(\boldsymbol{\theta}, \boldsymbol{Y})$, and its sampling variance can be obtained from $M (>1)$ sets of plausible values.

(a) Using each set of plausible values $\hat{\theta}_m$ in turn, evaluate t as if the plausible values were true values of θ. Denote the results $\hat{t}_m$, for $m = 1, \ldots, M$.

(b) Using the multiple weight jackknife approach (see Johnson & Rust, 1992; Johnson, Rust, & Hansen, 1990), compute for each m the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

(c) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \hat{t}_m / M.$$

(d) Compute the average sampling variance over the M sets of plausible values to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} U_m / M.$$

(e) Compute the variance among the M estimates $\hat{t}_m$ to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} (\hat{t}_m - t^*)^2 / (M - 1).$$

(f) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B_M.$$

Five sets of plausible values are used in NAEP analyses and are provided on the NAEP secondary-use data files for secondary analysis. Due to the excessive computation that would be required, NAEP analyses do not compute and average jackknife variances over all five sets of plausible values; they compute only the first set. Thus, in NAEP reports, U^* is approximated by U_1.

Statistical Tests

Suppose that, if θ values were observed for sampled students, the statistic $(t - T) / U^{1/2}$ would follow a t distribution with d degrees of freedom (see Johnson & Rust, 1992, on calculating degrees of freedom in NAEP sampling designs). Then the incomplete data statistic $(t^* - T) / V^{1/2}$ is approximately t distributed, with degrees of freedom given by

$$\nu = \frac{1}{\dfrac{f_M^2}{M - 1} + \dfrac{(1 - f_M)^2}{d}},$$

where f_M is the proportion of total variance due to the latent nature of θ:

$$f_M = (1 + M^{-1})\, B_M / V.$$

When B_M is small relative to V, the reference distribution for incomplete data statistics differs little from the reference distribution for the corresponding complete data statistics. This is the case with most major reporting variables in the main NAEP surveys. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each $\boldsymbol{U}_m$ and $\boldsymbol{U}^*$ is a covariance matrix, and $\boldsymbol{B}_M$ is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(\boldsymbol{T} - \boldsymbol{t}^*)\, \boldsymbol{V}^{-1} (\boldsymbol{T} - \boldsymbol{t}^*)^1$$

is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1 + M^{-1}) \text{ Trace } (\boldsymbol{B}_M \boldsymbol{V}^{-1}) / k.$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables $\mathbf{y}^c$ are consistent estimates of the corresponding population values T. Statistics involving background variables y, that were not conditioned on, or relationships among proficiencies from different content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the conditioning variables and to the proficiency of interest.

The direction of the bias is typically to underestimate the effect of nonconditioned variables. (For details and derivations, see Beaton & Johnson, 1987, 1990; Mislevy, 1991; Mislevy & Sheehan, 1987, sec. 10.3.5.) For a given statistic t^* involving one content area and one or more nonconditioned background variables, the magnitude of the bias is related to (a) the extent to which observed responses $\mathbf{x}$ account for the latent variable θ and (b) the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in all secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- High shared variance between conditioned and nonconditioned background variables mitigates biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions.
- High shared variance exacerbates biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

In the 1984 NAEP reading assessment, the magnitude of shrinkage for the subgroup means of a background variable that was not conditioned on averaged about 15%. Biases in multiple regressions that included conditioning variables averaged about 35%. Since that time, two important steps have been taken to greatly reduce potential biases of this type. First is the move to the focused-BIB matrix-sampling design, in which all the cognitive tasks a respondent is administered are drawn from the same subject area. On the average, respondents are presented about twice as many tasks in the subject area than would have been presented under the full BIB-spiraling design, which administered each examinee tasks from one, two, or three subject areas. This increases the extent to which $\mathbf{x}$ accounts for θ and, as noted above, decreases potential biases in all secondary analyses. Second is the increase in the number of background variables that can be included in the conditioning vector. This increases the number of secondary analyses that can be carried out with little or no bias and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Bruce Kaplan and Jennifer Nelson's analyses of 1988 reading data (personal communication, January 12, 1989; some results are summarized in Mislevy, 1991) indicate that these improvements have slashed the potential bias for nonconditioned variables in multiple regression analyses from the 1984 level of 35% to approximately 10% and have slashed biases in simple regression of such variables from 15% to 5%.

Table 1 gives representative results from an analysis in which Nelson estimated a number of substantively important effects from the 1988 NAEP reading data for 13 year olds in three ways: (a) with the operational conditioning process, which included the listed effects in the conditioning vector, so that estimates are consistent; (b) with no conditioning at all, so that the biases would be at their maxima; and (c) with conditioning on the first 32 principal components of the matrix of all 64 original conditioning vectors. The bias estimate for the male–female effect with no conditioning is calculated, as an example, as 100 (1 − 14.4/15.7). Note that the biases in analyses involving the original effects are virtually eliminated in the componential approach with only half as many conditioning variables.

A check on the impact of the approximations and simplifying assumptions employed in the ARM approach was carried out with data from the 1984 NAEP writing assessment (Beaton & Johnson, 1990). As a comparison for subgroup average writing scores, the same statistics were calculated using a totally different approach—the model-free, unbiased estimate for average responses based on the methodology employed by the Education Commis-

TABLE 1
Estimated effects based on full, no, and partial conditioning

	Conditioning				
Effect	Full*	None	Bias	32 components	Bias
Male–Female	−15.7	−14.4	−8%	−15.9	1%
White–Black	26.6	23.8	−11%	26.3	−1%
High metropolitan–low metropolitan	32.6	30.5	−6%	32.8	1%
Northeast–Southeast	10.4	9.4	−10%	10.1	−3%
13-year-old eighth graders–13-year-old seventh graders	32.7	29.2	−11%	32.7	0

* Imputations constructed with conditional distributions that included 64 contrasts, including those shown here.

sion of the States in previous NAEP analyses. This method is prohibitively expensive to be used for all NAEP statistics, but it could be calculated for the 44 questions in the common background questionnaire. Beaton and Johnson found that statistics based on the ARM were nearly indistinguishable from the model-free averages for those subgroups distinguished as conditioning variables and for subgroups whose memberships were well predicted by conditioning variables. Estimated standard errors were also smaller for the ARM estimates, because the ARM uses the available information about the relationships between exercises more efficiently. For those subgroups that were neither conditioned on nor well predicted by conditioning variables, the ARM exhibited biases. See Beaton and Johnson (1990) for an analytic treatment of the potential biases of the ARM under incomplete conditioning.

A Numerical Example

To illustrate how plausible values are used in analyses, this subsection gives some of the steps in the calculation of 1988 grade-level reading means and their estimation-error variances. The weighted mean of the first plausible values of the grade four students in the sample is 230.68, and the jackknife variance of these values is 1.17. Were these values true θ values, then 230.68 would be the estimate of the mean, and 1.17 would be the estimation-error variance. The weighted mean of the second plausible values of the same students, however, is 230.60; the third, fourth, and fifth plausible values give weighted means of 230.19, 230.32, and 230.06. Because all of these figures are based on precisely the same sample of students, the variation among them is due to uncertainty about the students' θs, having observed their item responses and background variables. Taking the jackknife variance estimate from the first plausible value, 1.17, as our estimate U^* of sampling variance and the variance among the five weighted means, .09, as our estimate B_M of uncertainty due to not observing θ, we obtain as the final estimate V of total error variance $1.17 + (1 + 5^{-1}).09 = 1.28$. With V and B_M defined as above, and with $M = 5$, we may approximate the proportion of variance in the estimated mean due to the latent nature of θ: $f_M = .08$. Corresponding values calculated for grade 8 and grade 12 are shown in Table 2.

Where Have We Come From, Where Are We Going?

IRT, with its claim of invariant item parameters, appeared to offer estimates of trends over time and comparisons between groups on a permanent scale of measurement, despite item sampling designs that presented different items to different students and changes in the item pool over time (Bock, Mislevy, & Woodson, 1982). IRT had proved its practical worth in the setting of individual measurement, solving previously intractable problems

TABLE 2

Estimation error variances and related coefficients for the 1988 grade-level reading assessments

Grade	U^*	B_M	V	f_M
4	1.17	.09	1.28	.08
8	.96	.06	1.03	.07
12	.69	.02	.71	.03

in equating, adaptive testing, and test construction. Many public discussions that took place during the time of the 1983 NAEP grant competition included considerations of some sort of IRT scaling.

As it turned out, existing IRT techniques fell short in the NAEP setting with its small numbers of items per student, complex sampling design, and need to measure very small differences and changes. In particular, the standard IRT expedient of using point estimates for individual examinee ability parameters as if they were the true parameters themselves yielded unacceptably biased estimates of population characteristics. Ignoring this source of uncertainty—and it is indeed usually ignorable when measuring and comparing individual students (Sheehan & Mislevy, 1988)—simply could not be tolerated. The IRT literature, having evolved in the setting of individual measurement where this problem did not arise, offered no way to integrate the uncertainty structure associated with individual measurement with the uncertainty structure associated with sampling students in a complex design. What was needed was a more comprehensive view of the nature of the data gathered in an assessment and models to capture the salient features of those data.

The general inferential framework is a synthesis emerging from recent developments in latent variable estimation (e.g., Andersen & Madsen, 1977, Mislevy, 1984, 1985), missing-data technology (Little & Rubin, 1987; Rubin, 1987), and hierarchical analysis (e.g., Bock, 1989). Ideally, a unified analytical framework for NAEP data would encompass all sources of information, all sources of uncertainty, and the joint and hierarchical relationships among them. The analysis would explicate the structure of variation among students, including both the background variables gleaned in the survey and the nesting and stratification variables used in sample selection. It would address variation among responses within students, including shared variation among items tapping related educational concepts.

Neither the computational capacity nor the armamentarium of procedures necessary to carry out this vision in detail is currently available for a project of the size and complexity of NAEP. Rubin supplied practical means of approximating the results of the theory he laid out in *Multiple Imputation* *for Missing Data in Sample Surveys* (Rubin, 1987). Viewing latent variables as missing data and employing randomization-based methods such as the jackknife to handle respondent sampling, one can employ his methods to draw justifiable inferences about the distributions of latent variables in settings such as NAEP's, and, in the process, one can provide secondary users with files of filled-in data sets to recover characteristics of population distributions that cannot be estimated consistently from traditional every-pupil test scores.

Although Rubin's work lays the theoretical foundation on which NAEP's scale-score analyses are based, practical implementation must be thoughtfully conceived, tested, revised, and continually improved and refined. Highlights of the chronology of multiple imputation methods in NAEP follow.

1985 *The Reading Report Card* (1985) introduced plausible values methods in an analysis of trends in reading over a decade of NAEP assessments for a small number of traditional NAEP reporting variables.

1985 To analyze reading proficiencies of language minority students (Baratz-Snowden & Duran, 1987), plausible values methods were extended to support analyses over a broader range of background survey variables.

1985 Beaton and Johnson introduced plausible values methodology for general linear models—the average response method (ARM)—for *The Writing Report Card* (Applebee, Langer, & Mullis, 1986).

1986 The multivariate extension of plausible values was introduced for the NAEP survey of young adult literacy (Kirsch & Jungeblut, 1986), providing for the joint analysis of four literacy score scales.

1987 Multivariate methods were enhanced for the analysis of the 1986 NAEP mathematics and science scales, using improved computing approximations to accommodate larger numbers of scales and fewer responses per student.

1989 Conditioning variables were introduced for school effects, leading to improved precision for secondary analyses using hierarchical models.

1990 Conditioning on principal components of background effects increased the precision of a broader range of secondary analyses (with nearly 200 background effects included in the process) while at the same time increasing the efficiency of the numerical approximations. This technique made it possible to use plausible values technology for the 40 reporting units in the 1990 NAEP Trial State Assessment of mathematics.

More work remains. Current NAEP research in the area of scaling concerns making approximations more accurate, extending the range of secondary analyses with negligible bias, incorporating variance terms for uncertainty associated with estimates of IRT parameters, and, of particular relevance in an era of increasing interest in performance assessment, extending the methodology to rating scaling and partial credit IRT models.

Note

[1] Muraki is investigating NAEP applications of multiple-category and ordered-response IRT models introduced by Bock (1972) and Samejima (1969). See Muraki (1990) on marginal estimation of item parameters in these models. The machinery of plausible values adapts immediately to ordered-response IRT models, differing from 3PL procedures only in the generation of likelihood functions for θ from observations.

References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42,* 357–374.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). *The writing report card: Writing achievement in American schools* (Report No. 19-W-02). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Baratz-Snowden, J. C., & Duran, R. (1987). *The educational progress of language minority students: Findings from the 1983–84 NAEP reading survey.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983–84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E. (1988). *Expanding the new design: The NAEP 1985–86 technical report* (No. 17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E., & Johnson, E. G. (1987). The average response method (ARM) of scaling. In A. E. Beaton, *The NAEP 1983–84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15,* 9–38.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985–86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data.* San Diego: Academic.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage in educational assessment. *Educational Researcher, 11(2),* 4–11, 16.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—Or should we? *Psychological Bulletin, 74,* 68–80.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B, 39,* 1–38.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: American Council on Education/Macmillan.

Johnson, E. G. (1990). Data analysis for the writing assessment. In E. G. Johnson & R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report* (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175–190.

Johnson, E. G., Rust, K. F., & Hansen, M. H. (1990). Weighting procedures and estimation of sampling variance. In E. G. Johnson & R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (NAEP Rep. No. 16-PL-02). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician, 37,* 218–220.

Little, R. J. A., & Rubin, D. B. (1987) *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton, *The NAEP 1983–84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing* (Research Rep. No. RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

National Assessment of Educational Progress. (1985). *The reading report card: Progress toward excellence in our schools* (Report No. 15-R-01). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34,* (4, Pt. 2).

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures* (Research Rep. No. 88-38-ONR). Princeton, NJ: Educational Testing Service.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika,, 36,* 427–436.

Authors

ROBERT J. MISLEVY is Principal Research Scientist, Educational Testing Service, Princeton, NJ 08541. He specializes in educational measurement large-scale assessment.

EUGENE G. JOHNSON is Coordinating Director, NAEP Research, Educational Testing Service, Princeton, NJ 08541. He specializes in statistical analysis, educational statistics, and design and analysis of sample surveys.

EIJI MURAKI is Research Scientist, Educational Testing Service, Princeton, NJ 08541. He specializes in item response theory and large-scale assessment.