DOCUMENT RESUME

ED 346 112 TM 018 330

AUTHOR Mazzeo, John; And Others
TITLE The Use of Collateral Information in Proficiency Estimation for the Trial State Assessment.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Apr 92
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Comparative Analysis; *Competence; Equations (Mathematics); *Estimation (Mathematics); *Factor Analysis; Grade 8; Junior High Schools; Junior High School Students; *Mathematical Models; Mathematics Tests; National Surveys; *Statistical Distributions
IDENTIFIERS Ability Estimates; Criterion Variables; National Assessment of Educational Progress; Plausibility Approach; *Proficiency Distributions; *Trial State Assessment (NAEP)

ABSTRACT

The adequacy of several approaches to estimation of proficiency distributions for the Trial State Assessment (TSA) in eighth grade mathematics of the National Assessment of Educational Progress was examined. These approaches are more restrictive than the estimation procedures originally used, with the same kind of plausible-values approach that has been implemented since 1984 with the national assessment. The proposed approaches are computationally less burdensome and could provide improved performance from a statistical standpoint. Results from six procedures including the operational procedure for 1990 are compared to results obtained by a criterion procedure for eight TSA jurisdictions. With some modifications, alternatives were generated by crossing the method of obtaining principal components, the method of selecting principal components, and the procedure for estimating the model. Methods that estimate a single population model did not provide acceptable results. Results support use of the less restricted models, particularly the one used operationally in 1990, producing separate sets of principal components for each jurisdiction from each jurisdiction's within-state correlation matrix, with decisions about the components to include made separately for each jurisdiction on the basis of the percent-of-trace criterion applied to the jurisdiction's correlation matrix. A separate population model is then estimated for each jurisdiction. Fourteen tables, 2 appendices presenting data from the analyses, and a 20-item list of references are included. (SLD)

THE USE OF COLLATERAL INFORMATION IN PROFICIENCY ESTIMATION FOR THE TRIAL STATE ASSESSMENT

John Mazzeo
Eugene Johnson
Drew Bowker
Y. Fai Fong

Educational Testing Service

Paper presented as part of the symposium: Statistical and Psychometric Issues in the Trial State Assessment, American Educational Research Association 1992 Annual Meeting.

THE USE OF COLLATERAL INFORMATION IN PROFICIENCY ESTIMATION FOR THE TRIAL STATE ASSESSMENT[1]

John Mazzeo
Eugene Johnson
Drew Bowker
Y. Fai Fong

Educational Testing Service

INTRODUCTION

Beginning in 1984, most of the results for the National Assessment of Educational Progress (NAEP) have been reported in terms of IRT-based scales (referred to here as *proficiency scales*). Estimates of features of the distribution of proficiencies (such as means, percentile locations, and proportions of students above certain levels) are routinely reported for important demographic subgroups, as well as for groups of students defined by their standing on a host of educational relevant variables.

In NAEP, the estimation of these proficiency distributions is based on a particular set of marginal estimation procedures sometimes referred to as the *plausible values approach* (Mislevy, Beaton, Kaplan & Sheehan, 1992) . As discussed in Mislevy (1991), the computing approximation used to carry out this approach is based on an extension of Rubin's (1987) multiple imputation procedures for survey nonresponse. Random draws, called plausible values in NAEP, are taken from predictive distributions, the parameters of which depend on students' responses to cognitive assessment items, other kinds of survey questions, and demographic variables. Following Rubin, multiple draws (five of them) are taken from the predictive distribution of each sampled examinee. When analyzed appropriately, these plausible values provide estimates of various features of the distribution of proficiencies, as well as appropriate estimates of the relationships between proficiencies and other variables, that are consistent under the model used to construct them.

The model-based predictive distributions which form the basis of the approach consists of two components, *a latent variable model*, and *a population model* (Mislevy, 1991). IRT provides the

---

[1]The work upon which this paper is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service. The authors are grateful to Dave Freund, Ed Kulick, Laura McCamley, for data processing and statistical programming. The authors would like to extend special thanks to Bob Mislevy for his always enlightening discussions on matters related to the current paper and more generally to NAEP scaling and analysis issues.

latent variable models used in NAEP. The population model, described further below, is similar in form to a multivariate regression model with IRT proficiencies being the dependent variable and demographic variables, survey variables, and, responses to other background questions forming the predictor variables. The parameters of both the latent variable and population models are estimated from the NAEP assessment data using marginal estimation procedures (see, e.g., Mislevy, Johnson, & Muraki, 1992).

In 1990, NAEP carried out its first Trial State Assessment (TSA) in eighth-grade mathematics. As with the national assessment, results were analyzed and reported using IRT-based scales for each of the 40 jurisdictions that agreed to participate. Separate scales were produced for five mathematics content areas. Results for each participant were estimated using the same kind of plausible-values based approach that has been implemented since 1984 with the national assessment. As applied to the TSA, the estimation procedures involved estimating a single latent variable model (the three-parameter logistic (3PL)), but separate population models (hence, separate sets of predictive distributions) for each of the participating jurisdictions (Mazzeo, 1991). The models were similar in form (i.e., linear multivariable regression models), however, the parameters of the model (i.e., the regression coefficients and residual variance matrices) were free to vary across jurisdictions.

The rationale for such an approach, discussed further below, was to ensure consistent (i.e., asymptotically unbiased) estimation of results for important subgroups within each of the jurisdictions. However, the approach was labor intensive, computationally intensive, and time consuming. With the planned expansion of the TSA to more academic subjects and more grades, along with desires for more timely reporting, a simpler, less labor intensive alternative producing comparable results would clearly be attractive. In addition, as discussed further below, simpler alternative procedures could, in principal, have more desirable statistical properties than the procedure actually used.

The research described here examines the adequacy of several more restrictive approaches within a plausible values framework. Because they are more restrictive, these approaches are computationally less burdensome and could provide improved performance from a statistical standpoint. Each of the procedures was used to reanalyze the results of the 1990 TSA for eight of the participating jurisdictions. Results obtained from each of the procedures and the procedure used operationally in 1990 are compared to results obtained by a criterion procedure (described below).

Overview of plausible values approach

Since its inception in the late 1960's, NAEP has used matrix sampling methods in an attempt to provide broad content coverage within the subject matter domains while maintaining acceptable limits on examinee testing time. Beginning in 1984, a particular variant of matrix sampling,

balanced-incomplete block spiraling, was coupled with Item Response Theory (IRT) scaling methods to provide a new approach the analysis and reporting of NAEP results (Messick, Beaton, & Lord, 1983). Since that time, NAEP results have, for the most part, been reported on scales derived from IRT.

IRT was developed in the context of measuring individuals, where each examinee is administered enough items to estimate his or her proficiency (θ) with a large degree of precision. In such cases, quantities that are of interest to consumers of NAEP reports (such as features of the distribution of θ for various demographic subgroups) are reasor ably approximated by distributions of point estimates of θ (such as maximum-likelihood estimates). However, this approach breaks down in the assessment setting where, due to limited testing time, each individual is administered relatively few items in a scaling area. The uncertainty associated with point estimates of θ is too large to ignore, and the features of the distribution of these estimates can be seriously biased as estimates of the distribution of θ (see, e.g., Mislevy, Beaton, Kaplan, & Sheehan, 1992).

In order to circumvent such difficulties, most NAEP results (such as proficiency means for subpopulations, the proportion of examinees above particular proficiency levels, and the relationships between proficiencies and educational variables) have been obtained using marginal estimation procedures which do not require point estimates for individual examinees. In NAEP, these marginal estimation procedures have been approximated using the so-called "plausible values" technology (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; Mislevy, Beaton, Kaplan, & Sheehan, 1992). The following is a brief overview of the plausible values approach as applied in the context of NAEP. A more thorough treatment can be found in Mislevy (1991).

In NAEP, samples of examinees are administered assessment instruments and background questionnaires and additional information is obtained from the teacher's of the sampled examinees as well as the principals of their schools. Thus, for each examinee, the observed data available are responses to a subset of cognitive items included in the assessment, as well as a variety of background questions relating to demographic characteristics and other educationally relevant variables. Let $\underline{y}$ be a vector of containing responses of all assessed examinees to all background variables, attitude questions, and survey design variables (such as school membership, or type of community). In other words, $\underline{y} = (\underline{y}_1,\underline{y}_2,..\underline{y}_n)'$, where $\underline{y}_i$ indicates the vector of background variables for the ith examinee. Similarly, let $\underline{x}=(\underline{x}_1,\underline{x}_2,...\underline{x}_n)'$ refer to the vector of responses of these students to the cognitive items included in the assessment. In addition, let $\underline{\Theta}^0=(\underline{\theta}_1,\underline{\theta}_2,...\underline{\theta}_n)'$ represent the vector of examinee proficiencies on the (possibly multiple) IRT scales of interest.

If the $\underline{\theta}_i$ were observed for each of the sampled examinees, it would be possible to compute a statistic $t(\underline{\Theta}^0,\underline{y})$ (e.g., a sample mean or sample percentile point for some subpopulation) to estimate some corresponding population quantity T. However, $\underline{\Theta}$ is unobservable. Following Rubin (1987), $\underline{\theta}_i$ is treated as "missing data" for all sampled examinees and $t(\underline{\Theta}^0,\underline{y})$ is approximated by its expected value given the observed data $(\underline{x},\underline{y})$,

$$t^*(\theta^0,y) = \mathcal{E}[t(\theta,y)|(x,y)] = \int_{-\infty}^{\infty} t(\theta,y)\, p(\theta|x,y)\, d\theta \qquad (1)$$

In NAEP, a Monte Carlo approximation to (1) is obtained by taking random draws from a predictive distribution of proficiencies for each examinee conditioned on their item responses and their responses to background questions and survey variables. These predictive distributions are denoted here as $p(\underline{\theta}|\underline{x},\underline{y})$. The values of these random draws are referred to as imputations in the sampling literature and plausible values in NAEP.

Of course, in NAEP, as in most applications, the predictive distributions are unknown. In order to evaluate the integral in (1), a particular model must be assumed and estimates, denoted $\hat{p}(\underline{\theta}|\underline{x},\underline{y})$, obtained. In NAEP, the predictive distributions are characterized as

$$p(\theta|x_i,y_i) = \frac{P(x_i|\theta)p(\theta|y_i)}{k} \qquad (2)$$

In (2), k^{-1} is a proportionality constant and $P(\underline{x}|\underline{\theta})$, what Mislevy (1991) refers to as the *latent variable model*, is the likelihood for $\underline{\theta}$ induced by the vector of responses to cognitive items $\underline{x}$ under an IRT model with conditional independence. The remaining piece, $p(\underline{\theta}|\underline{y})$, referred to by Mislevy (1991) as the *population model*, is the joint density of proficiency conditional on the observed value $\underline{y}$ of background responses. This joint density is assumed multivariate normal with mean given by $\Gamma'\underline{y}$, and covariance matrix Σ, where Γ is a matrix whose columns contain the coefficients for the regressions of each of the elements of $\underline{\theta}$ on $\underline{y}$. The model for the joint density of proficiency conditional on the background data is known in NAEP parlance as the *conditioning model*.

In NAEP, estimates of predictive distributions are obtained in two steps. First, assessment data are used to estimate IRT item parameter using BILOG (Mislevy and Bock, 1982). These item parameter estimates are then treated as known and used to fit a linear model to the assessment data of the form

$$\underline{\theta} = \Gamma'\underline{y}_i + \underline{\epsilon} \qquad (3)$$

where $\underline{\epsilon}$ is assumed normally distributed with mean zero and dispersion matrix Σ. Maximum likelihood estimates of Γ and Σ are obtained using Sheehan's (1985) MGROUP computer program. The program uses a variant of the EM solution described in Mislevy (1985) in which a normal approximation to $P(\underline{x}|\underline{\theta})$ is used (for further details see Johnson & Mislevy, 1991, or, Mislevy, Johnson, & Muraki, 1992). Based on the MLE's for Γ and Σ, and the normal approximation to $P(\underline{x}|\underline{\theta})$, MGROUP then also produces an estimate of the predictive distribution for subject i, $\hat{p}(\underline{\theta}|\underline{x},\underline{y})$, from which the plausible values are drawn.

The plausible values approach applied to the 1990 TSA

As described in Mazzeo (1991), a common latent variable model for each of five content area scales was estimated for use with all 40 jurisdictions participating in the TSA. Procedures for estimating the populations models were more complicated.

Plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), student attitudes toward mathematics, student behaviors both in and out of school (e.g., amount of TV watched daily, amount of mathematics homework each day), the type of math class being taken (e.g., algebra, or general eighth-grade mathematics), the amount of emphasis provided by the students' teachers to various topics included in the assessment, as well as a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended. Overall, relationships between proficiency and more than 50 variables, taken directly or derived from the student, teacher, and school questionnaires, or provided by Westat, were to be estimated and reported. When expressed in terms of contrast-coded main effects and interactions, this resulted in a total of 167 variables (see Koffler, 1991, Appendix C for a listing).

As described in Mislevy (1991), statistics that involve proficiencies and variables that are explicitly incorporated in the predictive model for $\underline{\theta}$ are consistent estimates of their corresponding population values. Statistics involving variables not included in the model are potentially subject to asymptotic biases. The degree of bias to be expected is a function of the complement of test reliability and the extent to which an omitted variable is independent of the variables included in the predictive model.

To avoid biases in reporting results and to minimize biases in secondary analyses, it was desirable to incorporate as many of the 167 contrasts into the predictive model as possible. The same background set of contrasts were included in the predictive model for all 40 Trial State Assessment participants. However, a decision was made to estimate population models separately for each of the 40 TSA participants. Estimating separate population models for each state was more complex than the simpler alternative of estimating a single model for all 40 states. However, it was felt that there were significant potential problems associated with the simpler approach to warrant the more complicated approach. The need for separate population models for each state can be understood by examining the potential problems associated with estimating a single common model.

Under the assumptions of the model, estimating a single model for all 40 states would produce consistent estimates of the means for subgroups for those contrasts that were explicitly included in the model. For example, since a Race/Ethnicity contrast was included for Asian Americans, a consistent estimate of the mean proficiency of the total group of Asian-American

students represented by those students who participated in the Trial State Assessment could be obtained from the single model. But TSA results were to be reported separately for each state and for subgroups within the states. Given this reporting structure, the single model approach is problematic because it will produce consistent estimates of the mean proficiency of subgroups within each state only if the magnitude of the effect associated with a particular contrast is identical across all 40 jurisdictions. Thus, the single model approach is tantamount to assuming there are no state-by-contrast interactions. This assumption appeared unnecessarily restrictive. The least restrictive approach, the one chosen, was to estimate separate population models for each jurisdiction.

Estimating separate models for each jurisdiction was not without problems, both practical and theoretical. First, a number of exact and near multicolinearities existed among these predictor variables within each of the jurisdictions. In a standard regression analysis (e.g. unweighted least squares or maximum likelihood estimation), estimation of regression coefficients in the presence of such multicollinearities often results in computing problems and numerical instabilities. The M-step of each cycle of M-GROUP's EM algorithm is carrying out a maximum-likelihood estimation of Γ based on sufficient statistics calculated in its preceding E-step. Hence, similar problems arise in MGROUP when one tries to obtain numerically stable estimates of Γ, and, consequently, Σ. Identifying these colinearities and removing them by selectively deleting variables for each jurisdictions would be a time consuming task. A more time efficient alternative was to transform the original predictor variables into a set of linearly independent variables by extracting principal components from their correlation matrix. Principal components, rather than the original variables, were used as the y variables in the population model.

Estimating a regression model by maximum-likelihood using a full set of principal components is equivalent to estimating a model in terms of the original predictors (see, e.g., Joliffe, 1986). To see why, let **Z** be an n-by-q matrix variables to be predicted, let **W** be an n-by-p matrix of standardized predictor variables, let **B** be a p-by-q matrix of regression coefficient, and let **E** be a matrix of residual terms. Also let **A** be the matrix of orthonormal eigenvectors of **W'W** and let **V** = **XA** be the matrix of principal components. Then,

$$Z = XB + E = XAA'B + E = VG + E \qquad (4)$$

where **G** = **A'B** is the matrix of regression coefficients for the principal component scores. Even when multicollinearities are present, the full set of components can be used to obtain an unbiased estimate $\hat{\mathbf{B}}=\mathbf{A}\hat{\mathbf{G}}$ that avoids computational problems. However, such an estimate may be subject to undesirably large degrees of sampling variability (Gunst & Mason, 1977)

An additional reason to worry about sampling instabilities in regression coefficient estimates is the large variables-to-observations ratio that would result from estimating separate predictive models for each jurisdiction. A typical sample size for the 1990 TSA participants was about 2,500. With 167 variables in the model, the ratio of variables to observations is about 15, which would not be considered large for a simple random sample. NAEP data are collected according to a

multistage sampling design which inflates standard errors of most statistics relative to their standard errors under a simple random sampling design. Thus, because of the presence of multicollinearities, as well as the high variable to observation ratio, it seemed desirable to attempt to reduce the dimensionality of each states predictor variables.

One way to obtain estimators of regression coefficients with smaller sampling variances under the fixed effect model is to delete principal components associated with the smallest eigenvalues (see, e.g., Belinfante & Coxe, 1986; Coxe, 1984; Gunst & Mason, 1977). Equivalently, one can set coefficients for those components to 0 in the model and obtain an estimate of G subject to those restrictions. If H is the matrix of regression coefficients for the principal components with appropriate rows set to zero, and $\hat{\mathbf{H}}$ is its estimate, then $\hat{\mathbf{B}}^{*}=\mathbf{A}\hat{\mathbf{H}}$ is the corresponding restricted estimated for the coefficients in terms of the original variables. $\hat{\mathbf{B}}^{*}$ has smaller sampling variability then $\hat{\mathbf{B}}$, but may biased to the degree that the restrictions imposed on G are not true. Analyses by Kaplan and Nelson (see Mislevy, 1991) on the 1988 NAEP Reading data suggested that a relatively small subset of principal components would capture almost all of the variance and most of the complex intercorrelations among the set of original variables and would reduce most of the potential bias for primary and secondary analyses of NAEP results.

Based on the above considerations, population models for each jurisdiction were obtained by including only a subset of the full complement of principal components as predictor variables. The principal components were produced separately for each jurisdiction. In other words, we obtained a separate eigenvector matrix $\mathbf{A}_s$ for each jurisdiction based on that jurisdictions cross-product matrix of standardized predictors ($\mathbf{W}_s$). The subset of principal components to include in the model was then done separately for each jurisdiction. In other words, the restrictions to be placed on G was decided on separately for each state. The number of principal components included in the population model for each jurisdiction was that number required to account for approximately 90 percent of the variance in the original contrast variables. Mislevy (1991), shows that this puts an upper bound of 10 percent on the potential bias for all analyses involving the original variables. Finally, population models were estimated separately for each jurisdiction based on only that jurisdiction's data. In other words, separate estimates of the matrix of residuals ($\hat{\mathbf{\Sigma}}_s$) and separate restricted estimates of the model coefficients $\hat{\mathbf{\Gamma}}_s=\mathbf{A}_s\hat{\mathbf{G}}_s$ were produced for each jurisdiction.

Some alternatives to the approach used

The procedure used to estimate predictive models for the 1990 TSA participants was developed in an attempt to: a) avoid unnecessary assumptions about the constancy of the relationships between background variables and proficiencies across jurisdictions, b) avoid computational difficulties due to multicollinearities, and, c) reduce the potentially large sampling variability associated with using a high dimensionality predictor variable. The resulting procedure

was, however, quite labor intensive and computationally burdensome. Because of the possible expansion of the scope of the TSA in the future, and the desire to shorten timelines for the analysis and reporting of results, there was considerable interest in considering less labor and less computationally burdensome procedures. If such procedures produced equivalent, or near equivalent results to the one used in 1990, consideration could be given to implementing these procedures in 1992 or in future NAEP assessments.

From a more statistical standpoint, there was interest in examining alternative procedures for reducing the dimensionality of the set of predictor variables. As mentioned above, deleting principal components with small eigenvalues is equivalent to using a restricted estimator for the coefficients of the population structure model which are possibly biased but may have smaller variance than an unrestricted estimator. If the restrictions imposed are correct (i.e., if the principal components set to 0 are, in fact, unrelated to the variable to be predicted), such a restricted estimate also has smaller mean squared error than the unrestricted estimator. However, this is not always the case. However, Lott (1973), for one, has provided examples in the context of principal components regression in which components with small eigenvalues have substantial relationships with the dependent variable. Eliminating these components from the regression equation can produced restricted estimates with smaller variance, but their associated bias results in larger mean squared error (MSE). A logical alternative approach would be to delete components with the smallest correlations with the dependent variable of interest. Keeping components with the strongest correlation with the dependent variable tends to result in restricted estimators with larger variance, but smaller MSE.

Based on the considerations discussed above, six alternative procedures for estimating population models were compared. With some modifications due to practical considerations, the alternatives were generated by crossing three factors: 1) method of obtaining principal components, 2) method of selecting principal components, and, 3) procedure for estimating the model.

Factors

Method of obtaining principal components (across-state vs within-state)

Two different methods of obtaining linear combinations of the original variables were compared. The first method, *across-state principal components*, consisted of obtaining the principal components of an *aggregate* correlation matrix of predictor variables. This aggregate matrix was based on all available data from each of the 40 jurisdictions that participated in the 1990 TSA. The total sample size was about 100,800 with the sample sizes for the jurisdictions ranging from 1326 to 2843, with a typical sample size of about 2,500. The aggregate correlation matrix was produced using a rescaled version of the sampling weights used for reporting each jurisdiction's 1990 results. The rescaling was carried out so that the sum of the weights for all jurisdictions were equal. The

intent of the rescaling was to produce a correlation matrix pertinent to a synthetic population containing approximately equal numbers of students from each jurisdiction.

The second method, *within-state principal components*, consisted of obtaining separate sets of principal components for each jurisdiction from their *within-state* correlation matrix. As with the previous method sampling weights were used in generating the matrix. This later method is the one that was used operationally for the 1990 TSA.

In the notation of the previous section, the across state principal components involved obtaining a single set of eigenvectors A for use with all states. The values of variables in the population model for state s were $V_{as}=W_{as}A$, where W_{as} is a matrix of predictor variables standardized in terms of the aggregate means and variances of the predictor variables. For the within-state component method, the values of the variables for the population model for state s were $V_{ws}=W_{ws}A$, where W_{ws} is a matrix of predictor variables standardized in terms of the within-state means and variances of the predictor variables.

Selecting principal components (% of trace vs r-squared criterion)

Two approaches to selecting the subset of principal components to include in the model were used. The first approach, the approach used operationally for the 1990 TSA, was based on a "percent-of-trace" criterion. The principal components were ranked (from high to low) in terms of their associated eigenvalues and the first *s* components were selected such that the sum of the eigenvalues for these *s* components was greater than or equal to 90 percent of the trace of the correlation matrix. Applying this criteria to the within state components produces a separate sets of restrictions on G for each state. Applying, the criteria to the across state components results for a common set of restrictions on G for each state.

The second approach consisted of two slightly different variations, each of which used an "r-square" criterion. The first variation, applied to the within-state principal components, is based on the procedure suggested by Lott (1973) in the context of principal components regression. Lott's procedure is to include in the model that set of components which maximizes the adjusted squared multiple correlation (Kerlinger & Pedhazur, 1973, page 283) with the dependent measure. Because principal components are uncorrelated by construction, the squared multiple correlation (and hence, the adjusted squared multiple correlation) between the dependent variable and any set of these principal components is simply the sum of the squared zero-order correlations of the components in that set. This feature makes Lott's procedure particularly simple to implement. The principal components are sorted and ranked (from high to low) on the basis of their zero-order correlation with the dependent variable. One then determines the rank order of the principal component at which the adjusted multiple-correlation is maximized and includes in the conditioning model that component and all components with lower rank orders.

In the context of the present study, the dependent variable for the population model is $\underline{\theta}$.

Because of the multiple-imputation procedures used in NAEP, accurate determination of the correlations between principal component scores and θ would require that predictive models be already available. So for practical reasons, correlations with a surrogate measure of proficiency were used to rank the principal components. The measure used was a logit transformation of the proportion of items answered correctly by each examinees. There were seven distinct test booklets used for the 1990 TSA. Each booklet contains items from each of the five content areas for which TSA scales were produced and the spiral administration procedure used in NAEP results in each book being administered to randomly equivalent samples within each of the participating jurisdictions. However, the booklets are not designed to be parallel in terms of level of difficulty. Therefore, the logit scores from each booklet were further standardized to have a mean of 0 and standard deviation of 1 in the aggregate sample of examinees that participated in the 1990 assessment.

Applying Lott's adjusted r-square procedure to the across-state principal components resulted in some problems. Because of the extremely large sample size on which the across-state approach is based (100,800 cases), the adjusted-r-square criterion suggested that virtually all principal components should be included. As discussed earlier, such an outcome is undesirable for practical reasons. Therefore, an alternative r-square approach was needed for selecting the appropriate subset of across-state principal components. An examination of multiple correlations revealed that keeping the 83 components with the largest zero-order correlations with logit scores resulted in a multiple r-square that was 99 percent as large as the multiple r-square obtainable using all principal components. Therefore, this set of 83 components was used for the across-state principal component methods.

As with the percent of trace methods, applying the r-square based criteria to the within state components produces a separate sets of restrictions on G for each state and applying, the criteria to the across state components results for a common set of restrictions on G for each state

Type of population model (within-state vs across-state)

Two approaches were evaluated. The first approach, equivalent to what was done operationally for the 1990 TSA, was to estimate separate population models for each state. In other words, separate estimates $\hat{\Gamma}_s$ and $\hat{\Sigma}_s$ were obtained for each of the jurisdictions. The second approach, involved estimating a single population model for use in all the states. In other words, a single set of estimates, $\hat{\Gamma}$ and $\hat{\Sigma}$ were produced.

The six alternative procedures

A total of six alternative procedures were obtained by combining the various levels of the factors described above. The six procedures were as follows:

1 - within-state pc's/within-state models/select by % of trace

This procedure (denoted as *wwt*) was the procedure used operationally for the 1990 TSA. It consisted of producing separate sets of principal components from each jurisdiction's within-state correlation matrix. Decisions about which components to include in the model were made separately for each jurisdiction on the basis of the percent-of-trace criterion applied to each jurisdiction's correlation matrix. A separate population model was then estimated for each jurisdiction.

2 - across-state pc's/within-state models/select by % of trace

This procedure (denoted as *awt*) used a single set of principal components for each of the jurisdictions derived from the aggregate correlation matrix. The set to be included was determined by applying the percent-of-trace criterion to the aggregate correlation matrix. A separate population model was then estimated for each jurisdiction. Note that, unlike the wwt procedure, the same set of principal components was included in each jurisdiction's model or, equivalently, the same set of restrictions on the regression coefficients was imposed. However, separate estimates of these restricted model coefficients were obtained for each jurisdiction.

3 - across-state pc's/across-state model/select by % of trace

This procedure (denoted as *aat*) used the same set of across-state principal components as used for the awt procedure. However, unlike the previous two methods, a single model was estimated and used for each of the jurisdictions. In other words, the same set restrictions was imposed and a single set of restricted coefficients estimated.

4 - within-state pc's/within-state model/select by adjusted r-square

This procedure (denoted as *wwr*) is identical to procedure 1 except that decisions about which components to include in the model were made on the basis of the adjusted r-squared criterion applied separately to each jurisdiction's correlation matrix.

5 - across-state pc's/within-state model/select by r-square

This procedure (denoted as *awr*) is identical to procedure 2 except that the set of pc's to be included was determined by applying the 99% r-square criterion to the aggregate correlation matrix.

6 - across-state pc's/across-state model/select by r-square

Denoted as *aar*, this procedure used the same set of across-state principal components as the previous method but a single model was estimated and used for each of the jurisdic s.

Methods 1 and 4 represent the least restrictive alternatives. Because estimation is carried out separately for each jurisdiction, the model being fit allows regression coefficients associated with any particular variable to differ across jurisdictions. Because the principal components are developed and selected separately for each jurisdiction, a unique set of linear constraints are being imposed on each jurisdiction's coefficients. Methods 1 and 4 differ in that, within each jurisdiction, they impose different sets of constraints on the parameters.

Methods 2 and 5 are slightly more restrictive. Again the estimation of separate models for each state allows for different regression coefficients for each state. However, the use of a single set of principal components results in the same linear constraints being imposed on the model for all of the states. Methods 2 and 5 differ in that, within each jurisdiction, they imply different sets of these common constraints.

Methods 3 and 6 represent the most restrictive models. Fitting a single population model for all participants implies an identical set of same regression coefficients for all participants and using a single set of principal components implies the same set of restrictions. Again, methods 3 and 6 differ only with respect to the particular linear constraints being imposed on the regression coefficients.

METHOD

Data

The data used in the study was from the 1990 Trial State Assessment. In order to keep work loads and costs at acceptable levels, a subset of eight of the 40 1990 participants were selected for the study. The jurisdictions included in the study evidenced a diversity of demographic profiles and exhibited a fairly wide range of performance on the assessment. The jurisdictions studied were California, the District of Columbia, Florida, Hawaii, New Jersey, North Dakota, Texas, and the Virgin Islands.

Procedure

Data from the 1990 TSA for each of eight jurisdictions were reanalyzed using one each of the procedures described above. Estimation was carried out using MGROUP. In estimating the models, the operational 1990 IRT item parameters were used and results were obtained for all five content area scales. Since one of the alternatives (wwt) was the actual procedure used operationally for the analysis and reporting of the 1990 TSA results, no additional analysis was required for this

alternative. The remaining six alternatives required a partial reanalysis of the 1990 data. For the methods which estimated a single population model for all jurisdictions (aat and aar), principal components were determined and the population model was estimated using all available data from the 1990 TSA (i.e., using data from all 40 participants). Data were weighted so that each jurisdiction contributed equally to the resulting estimates. The same principal components used for the single population methods were used with the awt and awr methods.

The major purpose of using the plausible values approach in NAEP is to ensure that consistent estimates of important subgroup differences, such as male/female and white/black differences can be obtained. Therefore, the procedures were compared by focussing primarily on the magnitude of estimated subgroup differences on the NAEP proficiency scales and estimates of the within-subgroup variability in proficiency. Comparisons were confined to two of five content area scale, Numbers and Operations(NO), and Data Analysis, Statistics, and Probability (DASP). The NO scale is the longest of the 1990 grade 8 scales and, with a typical examinee being administered about 20 items. The DASP scale is the shortest of the 1990 grade 8 scales with a typical examinee being administered 8 items. The importance of the conditioning model used depends on the amount of information available about individual proficiencies (Mislevy, Johnson, & Muraki, 1992). Consequently, the greatest differences between the methods is to be expected for the DASP scale and least for the NO scale. An additional legitimate criteria for comparing the procedures involves comparing them in light of estimates of the variability (due to sampling and other sources) associated with each.

Each of the above alternatives was evaluated by comparing its results to those obtained using a close approximation to a full maximum-likelihood solution under the least restrictive model. A full maximum-likelihood solution under the least restrictive model would entail separate coefficients for each jurisdiction for each of the predictor variables included in the population model. Such a model would result in unbiased estimates of the coefficients, although the error variances for the model coefficients could be somewhat larger than those obtained under a more restricted model. Within a principal components framework, such a model would be estimated by using separate sets of components for each jurisdiction, including all components with non-zero eigenvalues, and fitting the model separately within each jurisdiction. However, the inclusion of principal components associated with near zero eigenvalues often results in computational instabilities.

The criterion procedure actually used was an approximation to the full maximum-likelihood solution which avoided the computational instabilities. The procedure involved obtaining separate sets principal components for each jurisdiction and deleting principal components which showed evidence of computational instabilities. Such instabilities arise for the components with extremely small eigenvalues and are evident when one examines correlations among such components or with these components and other variables. Across the jurisdictions that we studied, such instabilities could be effectively eliminated by deleting the set of principal components with the smallest eigenvalues such that the sum of the eigenvalues in the set was about 1 percent of the trace of the

matrix of all eigenvalues. The remaining principal components were then used to estimate a separate population model for each jurisdiction.

Appendix 1 shows the total number of original variables, the number of exact colinearities (i.e., the number of 0 eigenvalues) and the number of principal components deleted due to computational instabilities. Appendix 2 shows total group and subgroup sample sizes for each of the eight jurisdictions evaluated in the study.

RESULTS

Table 1 shows the number of principal components included in the model for the methods that selected principal components separately for each state. The number selected (or, equivalently, the number of restrictions imposed) by the adjusted R^2 criteria was less than or equal to the number selected by the percent of trace criteria for all but one jurisdiction (North Dakota). For the methods using a single set of principal components for all states, the percent of trace criteria resulted in substantially more components being selected (94) than did the 99 percent of R^2 criteria (71).

Table 2 provides estimates of the mean and standard deviation of the aggregate population of students from the eight jurisdictions included in the study, as estimated by the criterion method and each of the six alternative methods. For the NO scale, estimates of the means and standard deviations are quite similar for all six methods. The largest difference from the criterion method occurred in the standard deviation estimate produced by the aar method. However, even this difference is only .4 scale points. For the DASP scale, results from the two methods using a single across-state population model depart some from the remaining methods. Both the aat and aar estimates of the mean of the aggregate population are about 1 point higher than those obtained with the remaining methods, while the standard deviations are almost 3 points lower. Apparently, even for a fairly course aggregate statistic like the overall mean and standard deviation of a collection a states, using a single population model seems to introduce some distortion in locations and unit of scale. This demonstrates that the particular conditioning model used is more important for cases with lower levels of information about individual proficiencies.

The means and standard deviations in Table 2 were used to linearly equate the scales produced by each of the methods to the scale of the criterion method. This was done so that subsequent comparisons could be evaluated in terms of changes in results over and above those related to differences in location and unit of scale. It should be pointed out, however, that such adjustments affect primarily the single population methods as little distortion was present for the remaining methods. All subsequent results are in terms of these "equated" scales.

Table 3 shows the means for each jurisdiction deviated from the national mean for the criterion method and for each of the alternative methods. In order to provide some framework for evaluating the size of the differences produced by the various methods, Table 4 shows the

differences between each alternative method and the criterion method divided by an estimate of the standard error of the criterion method result. In general, we regard small differences in standard error units as being below the threshold of noise inherent in the survey. It should be noted that the standard error shown reflect estimation error due to sampling examinees and error due to using multiple-imputation estimates of proficiency (Mislevy, Johnson, & Muraki, 1992). However, the production version of the MGROUP program available at the time of these analyses estimates the population model and draws multiple-imputations treating Γ estimates as fixed across imputations. As a result, uncertainty associated with estimating Γ is not reflected in these standard errors. (The program to be used for the 1992 analyses incorporates this source of variance).

For the NO scale, differences between the results from the criterion method and all four of the procedures that estimated separate population models for each jurisdiction were, with one exception, small. However, the estimate of the mean for the Virgin Islands (the most extreme jurisdiction) differed from the criterion method estimate by over two standard errors under the awt method and close to one standard error under the awr method. It is interesting to note that both methods slightly underestimated how far the Virgin Island mean was from the national mean. For the methods that estimate a single population model, results from the aar method differed little from the criterion method. The aat method showed slightly larger differences for two of the participants (Hawaii and DC) but all differences were less than a standard error.

Results are less satisfactory for the DASP scale. The means for Hawaii and D.C. differ from the criterion method by well over a standard error for both the methods which used across-state population models. The differences for the two states were in opposite directions however. Hawaii's distance from the national mean was slightly underestimated by both methods while D.C.'s was slightly overestimated.

Table 5 shows estimates of standard deviations for each jurisdiction obtained under the criterion method. Table 5 also contains for each of the alternative methods the ratio of its estimated standard deviation to the criterion method estimate. Table 6 provides a listing of the difference from the criterion method standard deviation divided by the standard error of the criterion method standard deviation. This standard error is the square root of the sampling variance of the criterion method standard deviation, estimated by the jackknife repeated-replications procedure (Johnson & Rust, 1992). For both the NO and DASP scales, results from the methods which estimate a single population model differ noticeably from the criterion method results. For the NO scale, differences exceed two standard errors for three of the eight jurisdiction for the aat method. Results for the aar method are only slightly better. Even more dramatic differences are evident for the DASP scale, where results differ from the criterion method in some cases by 4 to 6 standard errors. Of the remaining methods, results appear closest to the criterion method for the methods which selected principal components by the percent-of-trace criterion. Both the wwt and awt methods evidenced only 1 of 16 differences larger than a standard error. The wwr and aar methods, though better than the single population model methods, did result in several differences

exceeding a standard error for each of the scales.

As discussed earlier, a key area of interest in NAEP results involves examining differences in academic achievement among the various demographic groups, as well as among groups defined by their standing with respect to certain policy relevant educational variables. As shown in Mislevy, Beaton, Kaplan, & and Sheehan (1992), the predictive models used in NAEP are essential to ensure accurate estimation of subgroup differences when examinees are administered a relatively small number of items. Therefore, it is of particular interest to examine estimates of subgroup differences obtained under each of the alternative methods. In addition, it of equal interest to examine the extent to which the number of items taken by each examinee mitigates the effects of differences in the methods of obtaining population models. As mentioned earlier, one would expect larger differences due to method for the DASP scale.

Table 7 contains estimates of male-female differences as estimated by each of the population methods. Table 8 contains the differences between each alternative method and the criterion method divided by the criterion method standard error. Differences between the alternative methods and the criterion method are quite small for the NO scale, with one notable exception. Under the criterion method, the gender difference in Hawaii was estimated as 8 points favoring females, compared to 0 points for the nation as whole. Under the aat and aar method, Hawaii's gender difference is shrunk to 5 and 4.3 points respectively, differences that are on the order of 2 standard errors. Hawaii's results are equally unsatisfactory under the across-state methods for the DASP scale. In addition, the results for the Virgin Islands suggest that the size of the gender difference is underestimated, relative to the criterion method, by all six alternative methods and the underestimation is fairly substantial for three of methods (awt, wwr, and aar). Somewhat surprisingly, one of the three methods showing substantial differences from the criterion method was the wwr method, which involves estimating separate models for each state.

Table 9 contains estimates of mean proficiency differences between white and hispanic students and Table 10 shows the differences in terms of criterion method standard error units. For both the NO and DASP scales, it is apparent the methods using a single population model perform markedly less well than the other methods. As might be expected from the earlier discussion, the across-state methods performed particularly poorly for the DASP scales. In Florida, where the difference is somewhat smaller than that observed nationally, and for New Jersey and Texas, where the difference is somewhat larger than that observed nationally, white-hispanic differences are substantially overestimated by both the single population model approaches. By contrast, in D.C. and Hawaii, where the differences are larger than that observed nationally, the white-hispanic difference is substantially underestimated.

As part of the NAEP survey, the teachers of the assessed students were asked a variety of questions about their classes. One such question, included as a population variable, asked teachers to indicate the ability level of their classes (high, medium, low, or mixed). Table 10 shows mean proficiency differences between students in high and low ability classes, as identified by their

teachers. Table 11 contains estimates of mean proficiency differences between students in high and low ability classes and Table 12 shows the differences in criterion method standard error units. For Hawaii, where the difference was much larger than the difference observed nationally, results from the across-state population models badly underestimated the size of difference between high and low ability classes. This underestimation occurred for the NO scale but was particularly noticeable for the DASP scale where the difference was estimated under the single population methods was only about half as large as that estimated under the criterion method.

As part of the NAEP survey, the assessed students were asked a variety of questions about their backgrounds, study habits, and out-of-school activities. One such question, included as a population variable, asked students how much television they watched each day (1 hour or less, 2 hours, 3 hours, 4-5 hours, 6 or more hours). Table 13 shows mean proficiency differences between students watching 1 hour or less and students watching 6 or more hours and Table 14 shows differences from the criterion method in standard error units. With one exception, differences between each of the methods and the criterion method are small for the NO scale. However in the Virgin Islands, where the difference slightly favors students watching 6 or more hours of TV, the across-state population procedures reverse the direction of the estimated effect. This effect is even more dramatically evident for the DASP scale. Based on the criterion procedure, the mean proficiency for Virgin Island's students reporting more than 6 hours of TV watching is almost 11 point higher than the proficiency mean for students reporting 1 hour or less of TV watching. This effect is markedly different than that observed for the remaining jurisdictions and for the nation as a whole. Perhaps this is due to the demographics of Virgin Islands. Students reporting less than 1 hour of TV watching may come from lower SES homes where TV's may be less prevalent. In contrast, the Virgin Island's results based on the across-state population methods suggest that students with 1 hour or less of TV watching have a 2 to 3 point advantage of students reporting 6 or more hours of TV watching. Apparently, the estimated difference has been shrunk dramatically toward the national difference.

SUMMARY AND CONCLUSIONS

For both practical and theoretical reasons, the study reported here examined the adequacy of alternative procedures to estimating population models for use in NAEP's multiple imputations procedures. From a statistical viewpoint, the alternative procedures can be considered as methods for obtaining estimates of the coefficients of the model subject to sets of linear constraints. The restricted estimators, though biased, may have superior properties from the point of view sampling variability or MSE. The least restrictive procedures (wwt and wwr) allow for separate estimates of Δ and impose separate sets of linear constraints for each jurisdiction. Slightly more restrictive procedures (awt and awr) allow for separate estimates of Γ for each jurisdiction but impose a common set of constraints on all such estimates. The most restrictive procedures (aat and aar)

require a single estimate of Γ and a common set of constraints. The methods of choosing principal components (percent-of-trace based and R^2-based) can be viewed as different empirical procedures for identifying the constraints to be imposed.

From a practical viewpoint, the more restricted alternatives can be viewed as ways of attempting to reduce the work load and computing requirements associated with the TSA, hence shortening reporting deadlines and reducing costs. Using the 1990 TSA as an example, the wwt and wwr method involve 40 sets of analyses to determine principal components and 40 sets of analyses to carry out model estimation. The awt and awr methods still require 40 sets of model estimation analyses, but the number of principal component analyses is reduced to 1. The aat and aar methods require only a single set of principal component analyses and a single set of model estimation runs. It should be clear from a practical standpoint why, if acceptable results could be obtained, the restricted procedures would be attractive.

In our opinion, the analyses reported here indicate that the methods which estimate a single population model *do not provide acceptable results*. Even for highly aggregated quantities such as the mean and standard deviation of the composite population of the eight states studied here, the aat and aar methods did not produce adequate results. When compared to the criterion procedure, the aat and aar methods showed substantial differences in estimating the mean and standard deviations for several of the eight jurisdictions studied here. For each of the four empirical contrasts examined (male-female, white-hispanic, high ability-low ability, and 1 hour of TV-6 hours of TV), results from the single-population-model methods differed noticeably from the results obtained from the criterion method for several of the jurisdictions. In particular, constrasts the that were markedly different in magnitude than those observed for the nation (e.g., ability group differences in Hawaii) were often badly underestimated. These results suggest that the relationship between background variables and proficiencies is sufficiently different across jurisdictions necessitate the use of separate population models for each TSA participant.

Differences in performance among the remaining alternatives are less distinct. However, on balance, one would have to conclude from the results presented here that the wwt method, the method used operationally for the 1990 TSA, produced the most satisfactory results. Across all of the state mean and contrast comparisons carried out, results from the wwt method differed by less than a standard error from those obtained using the criterion method. In most cases, the differences were considerably less. The single instance in which results for the wwt method differed from those obtained using the criterion method occurred for the standard deviation of Texas on the DASP scale (1.04 standard error units). Consequently, current plans are to continue to use the wwt method for analysis of the 1992 TSA.

While the remaining separate-model methods did not perform poorly, their results were clearly less satisfactory than the wwt method. Differences in overall state means, and estimates of empirical contrasts were generally small for the wwr method. For only two of the contrasts did the differences from the criterion method exceed one standard error. In both cases, this occurred on

the DASP scale. However, in estimating state standard deviations, the wwr method performed noticeably less well than the wwt method. For the former, five of differences from the criterion method estimates exceeded a standard error in magnitude. The corresponding number was only 1. The awt and awr methods did not often result in large differences from the criterion method. However, a few substantial differences in overall state and subgroup means were found, though almost all appeared to be limited to the awt method applied to the data from the Virgin Islands. For the aar method, state standard deviations were misestimated somewhat for Hawaii and North Dakota.

Although the results reported here tend to support the use of the less restricted models in general, and the wwt method in particular, a definitive statement on the various methods would be somewhat premature. The analyses reported on here have focussed on comparing *point estimates* of state means, state standard deviations, and selected empirical contrasts. Such comparisons are extremely important from the point of view of identifying unacceptable biases that might result from the various procedures. In our judgement, their is sufficient evidence of unwanted biases in the single population models to warrant their exclusion from future consideration. However, results are less clear cut for the various procedures which estimate separate population models for each state and, before making definitive conclusions, it would probably be wise to examine these alternatives from the point of view of *interval estimates*, sampling variances, and MSE. Future studies will extend this work along those lines.

In future research, we also intend to expand the list of alternative procedures somewhat. For example, one procedure being given consideration involves combining contrasts and principal components in a single model. The method would proceed as follows. First a relatively small set of key variables would be identified. Most likely, this set of variables would consist of the major NAEP reporting variables (gender, race/ethnicity, type of community, parental education, # of reading materials in the home, etc.). Contrasts would be defined for these variables. The remaining larger set of background variables would also be expressed as contrasts. The second set would be residualized for the first set and then subjected to a principal components analysis. The final set of variables appearing in the population model would consist of the first set of contrasts and some (hopefully) small number of principal components from the remaining contrasts. Preliminary analyses with this approach are encouraging (Nelson, 1992). However, the necessary evaluative research will most likely not be completed in time to impact the current assessment's results.

Before concluding, there are several additional points that should be noted regarding the quantification of uncertainty in NAEP results. The standard errors used in this report reflect two sources of error, a component due to sampling and a component due to imputing rather than observing θ. This later component is estimated by generating multiple plausible values for each examinee (NAEP uses 5), treating these as making up five full data sets, and quantifying the between set variance of results. Such an approach follows Rubin's (1987) suggestions for

incorporating imputation error. However, in NAEP's computing machinery prior to 1992 (MGROUP), each set of plausible values are selected from a predictive distribution calculated using a fixed set of item parameters and a fixed estimate of Γ. Thus, the uncertainty associated with these estimates in not reflected in the imputations component of variance. As a result, NAEP's reported standard errors are somewhat optimistic.

A more satisfactory approach, one consistent with Rubin's suggestions, would be to: 1) take a 5 random draws from the posterior distribution of item parameters, 2) take 5 random draws from the posterior distribution of Γ, and, 3) produce five sets of plausible values, each using a different estimate of item parameters and Γ. Such a procedure would provide more realistic estimates of the component of variability due to imputation. Plans for incorporating uncertainty about Γ are already in progress. Neal Thomas at ETS is producing enhanced versions of the MGROUP program that will accomplish this and plans are that standard errors for the 1992 NAEP results will reflect the uncertainty involved in using estimates of Γ. Work in progress by Mislevy, Sheehan, and Wingersky (1991) is directed toward incorporating information about the uncertainty arising from using item parameter estimates.

An additional source of uncertainty in NAEP results, one demonstrated here, is their sensitivity to models of nonresponse. As noted earlier, the plausible values approach used in NAEP is an adaptation of Rubin's (1987) model-based procedures for nonresponse in surveys. Rubin has talked about the need to display the sensitivity of results to different models of nonresponse.

> If in a particular survey without follow-up response, there is no single accepted class of assumptions about nonresponse, then it is obviously prudent to perform data analyses under a variety of plausible models for nonresponse. *If (a) inferences vary in important ways as the models change and (b) the data cannot eliminate some models as inappropriate, then the tautological conclusion must be that the data cannot support sharp inferences without further specification of the models*. (page 17, emphasis added)

In the current study, the various approaches constitute different models. While it appears to be the case that the data *can* eliminate some of the models (i.e., those which assume a single population model for all jurisdictions), a decision among the remaining models is less clear cut and one could perhaps consider then as competing alternatives.

With the exception of the single population approaches, differences in results among the alternative procedures did occur. Such differences were, for the most part small, and in our opinion unlikely to affect the types of inferences typically made about NAEP results. Nevertheless, such differences do underscore the fact that there is an additional source of uncertainty not reflected in standard errors for NAEP results, the uncertainty associated with choice of nonresponse model. Perhaps a direction for future work would be to consider how to incorporate this kind of uncertainty into published standard error estimates.

References

Belinfante, A. & Coxe, K. (1986) Principal components regression - Selection rules and application. *Proc. Bus. & Econ. Sec., Amer. Stat. Assoc.* 429-431.

Coxe, K. (1982) Multicollinearity, pricipal components regression, & selection rules for these components. *Proc. Bus. & Econ. Sec., Amer. Stat. Assoc*, 449-453.

Gunst, R.F. & Mason, R.L. (1977). Biased estimation in regression: An evaluation using mean squared error. *Journal of the American Statistical Association, 72*, 616-628.

Johnson, E.J. & Mislevy, R.J. (1991) Theoretical background and philosophy of NAEP scaling procedures. In S.L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment Program* (No. 21-ST01). Wash. DC: National Center for Education Statistics.

Johnson, E.J. & Rust, K.F. (1992) Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics* (in press).

Joliffe, I.T. (1986) *Principal component analysis*. New York: Springer-Verlag

S.L. Koffler (1991) *The technical report of NAEP's 1990 Trial State Assessment Program* (No. 21-ST01). Wash. DC: National Center for Education Statistics.

Kerlinger, F.N. & Pedhazur, E.J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston

Lott, W.F. (1973). The optimal set of principal component restrictions on a least-squares regression. *Communications in Statistics*, 2, 449-464.

Mazzeo, J. (1991) Data analysis and scaling. In S.L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment Program* (No. 21-ST01). Wash. DC: National Center for Education Statistics.

Messick, S., Beaton, A.E., & Lord, F. (1983). *A new design for a new era*. NAEP report 83-1. Princeton, NJ: National Assessment of Educational Progress.

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., Beaton, A.E., Kaplan, B. & Sheehan, K.M.. (1992) Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, (in press).

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., Johnson, E.J. & Muraki, E.J. (1992). Scaling procedures in the National Assessment of Educational Progress. *Journal of Educational Statistics*, (in press).

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1991) How to equate tests with little or no data. Unpublished manuscript.

Nelson, J. & Bowker, D. (1992, April) Using principal components in conditioning NAEP cross-sectional scales. Paper presented at the annual meeting of the American Educational Research Association.

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

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

Table 1 Sample sizes and number of principal components included in the population model for each state for the criterion, wwr and wwt methods

	N	crit	wwt	wwr
CA	2424	126	90	90
DC	2135	119	87	84
FL	2534	125	91	81
HI	2551	123	88	88
NJ	2710	122	89	81
ND	2485	117	86	100
TX	2542	124	90	84
VI	1326	115	82	78

Table 2 - Means and standard deviations for the aggregate population of the eight states estimated by each alternative procedure

NO Scale:

	crit	wwt	awt	aat	wwr	awr	aar
mean	257.8	257.8	258.0	257.8	257.8	257.9	257.9
sd	38.1	38.1	37.9	37.9	37.9	38.1	37.7

DASP Scale:

	crit	wwt	awt	aat	wwr	awr	aar
mean	247.4	247.6	247.1	248.7	247.6	247.5	249.0
sd	49.6	49.6	50.1	46.8	49.3	49.6	46.2

Table 3 - Adjusted state means deviated from national mean

NO Scale (National mean = 266)

	crit	wwt	awt	aat	wwr	awr	aar
ND	20.1	19.8	19.9	20.0	20.1	20.2	20.4
NJ	7.3	7.5	7.1	7.4	7.4	7.3	7.2
TX	-4.1	-4.1	-4.3	-4.1	-4.0	-4.2	-4.2
FL	-5.9	-5.8	-6.4	-6.1	-5.9	-6.1	-6.0
CA	-6.5	-6.5	-6.8	-6.4	-6.5	-6.7	-6.5
HI	-10.0	-9.9	-10.0	-9.3	-9.8	-10.0	-9.5
DC	-28.2	-27.8	-28.3	-28.8	-28.2	-28.4	-28.5
VI	-38.7	-39.0	-37.3	-38.6	-39.0	-38.1	-38.8

DASP Scale (National mean = 262):

	crit	wwt	awt	aat	wwr	awr	aar
ND	23.4	23.6	23.1	24.3	23.5	23.7	24.8
NJ	8.0	7.8	8.3	7.3	8.0	7.7	7.0
TX	-5.5	-5.8	-5.5	-5.8	-5.8	-5.6	-6.3
FL	-7.2	-7.5	-7.3	-7.9	-7.2	-7.3	-7.3
CA	-7.8	-7.8	-6.9	-7.9	-7.7	-7.5	-7.9
HI	-19.3	-19.8	-19.4	-17.4	-19.7	-19.6	-17.7
DC	-41.0	-40.6	-40.3	-42.2	-40.8	-40.7	-42.3
VI	-67.3	-66.5	-68.7	-67.1	-66.9	-67.3	-66.9

Table 4 Adjusted state means -- Deviation from criterion results in se(criterion) units

NO Scale:

	wwt	awt	aat	wwr	awr	aar
ND	-0.17	-0.15	-0.08	-0.01	0.05	0.18
NJ	0.13	-0.16	0.10	0.10	-0.04	-0.05
TX	-0.03	-0.14	-0.03	0.06	-0.06	-0.07
FL	0.05	-0.40	-0.20	-0.03	-0.14	-0.08
CA	-0.03	-0.21	0.04	-0.03	-0.13	-0.01
HI	0.07	0.00	0.83	0.27	0.02	0.51
DC	0.47	-0.13	-0.81	-0.05	-0.26	-0.42
VI	-0.52	2.40	0.18	-0.49	0.96	-0.23

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
ND	0.11	-0.17	0.54	0.05	0.15	0.81
NJ	-0.17	0.24	-0.47	-0.01	-0.25	-0.75
TX	-0.20	-0.02	-0.18	-0.18	-0.09	-0.44
FL	-0.23	-0.08	-0.49	-0.02	-0.10	-0.08
CA	-0.02	0.46	-0.07	0.04	0.13	-0.03
HI	-0.45	-0.11	1.93	-0.41	-0.25	1.56
DC	0.48	0.76	-1.32	0.27	0.39	-1.48
VI	0.90	-1.51	0.20	0.40	0.05	0.42

Table 5 Adjusted state standard deviations relative to criterion results from the criterion procedure

NO Scale:

	crit	wwts	awt	aat	wwr	awr	aar
CA	37.1	1.01	1.02	0.99	1.02	0.99	0.99
DC	31.9	1.00	1.01	1.05	0.97	0.98	1.02
FL	34.7	0.99	1.00	1.00	1.00	0.99	1.01
HI	38.2	1.00	1.01	0.93	1.02	1.03	0.95
NJ	35.2	0.99	1.00	1.04	1.00	1.01	1.04
ND	30.2	1.01	0.98	1.00	0.97	0.99	1.02
TX	34.0	0.99	1.00	0.98	1.01	1.00	0.99
VI	29.3	1.01	1.01	1.01	0.98	1.01	0.99

DASP Scale:

	crit	wwts	awt	aat	wwr	awr	aar
CA	43.9	0.99	0.99	1.00	1.00	1.00	1.01
DC	42.3	1.00	1.00	1.00	0.99	1.00	1.01
FL	42.5	1.00	1.00	1.04	1.02	1.01	1.04
HI	48.1	1.02	1.00	0.89	1.02	1.01	0.91
NJ	40.8	1.00	0.99	1.13	1.01	1.00	1.11
ND	33.2	0.99	0.97	1.05	0.98	0.93	1.05
TX	41.7	1.02	1.00	1.02	1.01	1.01	1.01
VI	41.7	1.00	1.01	0.89	0.97	1.01	0.88

Table 6 Adjusted state standard deviations - differences from criterion method results in se(criterion) units

NO Scale:

	wwt	awt	aat	wwr	awr	aar
CA	0.48	1.06	-0.42	0.74	-0.56	-0.54
DC	-0.17	0.53	2.10	-1.16	-0.88	0.99
FL	-0.30	0.14	0.08	-0.18	-0.45	0.34
HI	0.22	0.55	-3.25	1.00	1.24	-2.60
NJ	-0.43	-0.12	2.07	0.09	0.60	1.94
ND	0.32	-0.61	0.05	-1.12	-0.18	0.60
TX	-0.40	-0.18	-0.96	0.46	0.27	-0.80
VI	0.29	0.60	0.38	-0.75	0.40	-0.65

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
CA	-0.37	-0.50	-0.20	0.20	-0.10	0.55
DC	0.04	-0.10	0.03	-0.38	-0.12	0.41
FL	0.04	0.14	2.03	1.00	0.75	1.91
HI	0.85	-0.16	-6.03	0.96	0.46	-5.05
NJ	0.03	-0.40	4.60	0.33	0.17	3.71
ND	-0.24	-0.80	1.42	-0.69	-2.05	1.52
TX	1.04	-0.18	0.95	0.29	0.31	0.61
VI	-0.06	0.57	-4.14	-1.13	0.56	-4.81

Table 7 Male - female differences--deviated from national mean

NO Scale (National mean = 0):

	crit	wwt	awt	aat	wwr	awr	aar
ND	5.7	5.9	4.8	4.1	4.7	5.7	5.5
TX	4.0	3.4	3.9	3.3	4.3	3.7	3.3
NJ	3.5	3.7	3.4	2.8	3.5	3.0	3.0
FL	1.2	0.7	1.4	1.8	1.5	0.5	1.7
CA	1.0	0.8	0.5	2.2	1.1	0.1	1.6
VI	0.6	1.4	-0.2	1.1	0.4	0.2	0.6
DC	-3.6	-4.3	-4.3	-2.5	-3.2	-4.2	-3.3
HI	-8.0	-7.4	-7.3	-5.0	-8.3	-7.5	-4.3

DASP Scale (National mean =1):

	crit	wwt	awt	aat	wwr	awr	aar
ND	6.8	7.8	7.6	6.3	6.7	7.9	7.5
VI	5.7	4.3	2.4	4.2	1.9	4.4	1.1
FL	4.5	4.9	4.7	4.7	3.7	4.3	4.0
NJ	4.0	3.8	4.0	4.6	4.0	3.3	4.9
CA	3.0	3.5	2.9	3.6	1.9	3.7	3.9
TX	2.6	1.6	2.2	3.6	2.3	2.3	3.2
DC	-1.9	-2.5	-1.6	-1.6	-1.5	-2.5	-2.3
HI	-5.6	-5.2	-4.1	-2.2	-6.7	-5.3	-2.1

Table 8 M-F difference --Deviation from criterion method results in se(criterion) units

NO Scale:

	wwt	awt	aat	wwr	awr	aar
ND	0.08	-0.38	-0.67	-0.42	0.00	-0.08
TX	-0.29	-0.05	-0.33	0.14	-0.14	-0.33
NJ	0.10	-0.05	-0.35	0.00	-0.25	-0.25
FL	-0.24	0.10	0.29	0.14	-0.33	0.24
CA	-0.09	-0.22	0.52	0.04	-0.39	0.26
VI	0.53	-0.53	0.33	-0.13	-0.27	0.00
DC	-0.50	-0.50	0.79	0.29	-0.43	0.21
HI	0.35	0.41	1.76	-0.18	0.29	2.18

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
ND	0.37	0.30	-0.19	-0.04	0.41	0.26
VI	-0.58	-1.38	-0.63	-1.58	-0.54	-1.92
FL	0.15	0.08	0.08	-0.31	-0.08	-0.19
NJ	-0.09	0.00	0.26	0.00	-0.30	0.39
CA	0.17	-0.03	0.21	-0.38	0.24	0.31
TX	-0.36	-0.14	0.36	-0.11	-0.11	0.21
DC	-0.32	0.16	0.16	0.21	-0.32	-0.21
HI	0.20	0.75	1.70	-0.55	0.15	1.75

Table 9 White-Hispanic difference deviated from national mean[2]

NO Scale (National mean = 25):

	crit	wwt	awt	aat	wwr	awr	aar
DC	41.3	38.4	38.8	36.0	39.6	40.0	33.4
CA	8.0	7.8	8.7	7.8	8.0	7.8	8.5
ND	6.8	5.8	7.3	2.9	7.8	6.4	6.0
NJ	6.7	6.3	7.4	9.6	7.5	8.0	10.4
HI	5.3	5.4	5.6	2.9	6.9	6.5	2.8
TX	1.6	1.5	1.5	2.3	2.2	2.5	2.6
FL	-6.8	-7.0	-7.3	-5.8	-7.4	-6.4	-4.6

DASP Scale (National mean = 23):

DC	52.4	50.6	49.1	42.8	54.2	51.8	44.5
HI	13.7	13.7	12.6	4.4	11.8	10.6	6.6
CA	12.0	10.9	11.4	12.5	10.8	11.7	12.7
NJ	11.8	13.7	11.4	18.8	11.9	13.0	17.9
TX	2.0	2.4	1.7	6.2	2.1	3.5	6.3
ND	-3.9	-7.4	-4.9	-5.1	-0.6	-4.9	-2.1
FL	-9.8	-9.4	-10.9	-4.3	-8.6	-9.1	-3.4

[2]Results not reported for Virgin Islands due to insufficient sample sizes for white examinees

Table 10 White-Hispanic difference --Deviation from criterion method results in se(criterion) units[3]

NO Scale:

	wwt	awt	aat	wwr	awr	aar
DC	-0.57	-0.49	-1.04	-0.33	-0.25	-1.55
CA	-0.08	0.27	-0.08	0.00	-0.08	0.19
ND	-0.15	0.07	-0.57	0.15	-0.06	-0.12
NJ	-0.13	0.23	0.97	0.27	0.43	1.23
HI	0.03	0.09	-0.69	0.46	0.34	-0.71
TX	-0.05	-0.05	0.33	0.29	0.43	0.48
FL	-0.07	-0.19	0.37	-0.22	0.15	0.81
VI	-0.53	0.47	-0.24	0.17	0.36	0.08

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
DC	-0.19	-0.35	-1.01	0.19	-0.06	-0.83
HI	0.00	-0.27	-2.32	-0.47	-0.77	-1.78
CA	-0.33	-0.18	0.15	-0.36	-0.09	0.21
NJ	0.46	-0.10	1.71	0.02	0.29	1.49
TX	0.13	-0.10	1.40	0.03	0.50	1.43
ND	-0.51	-0.15	-0.18	0.49	-0.15	0.26
FL	0.13	-0.37	1.83	0.40	0.23	2.13
VI	-0.40	-1.01	0.67	1.28	0.54	1.35

[3]Results not reported for Virgin Islands due to insufficent sample sizes for white examinees.

Table 11 High-Low ability difference deviated from national mean

NO Scale (National mean = 49):

	crit	wwt	awt	aat	wwr	awr	aar
HI	19.2	19.5	19.5	12.4	19.6	19.1	12.7
FL	10.6	10.5	10.7	11.1	10.1	9.4	10.8
CA	10.4	9.2	10.2	8.7	10.1	9.6	8.5
NJ	9.1	8.3	8.6	10.7	9.5	8.4	10.2
ND	6.9	5.8	3.9	3.9	5.0	7.8	6.2
TX	5.5	5.8	6.0	5.2	3.9	5.6	4.5
DC	-4.3	-5.4	-4.9	-1.3	-6.1	-6.8	-2.5
VI	-14.7	-15.8	-14.4	-12.1	-14.8	-14.7	-11.6

DASP Scale (National mean = 57):

	crit	wwt	awt	aat	wwr	awr	aar
HI	30.0	31.3	29.2	16.0	29.1	28.9	16.8
FL	13.7	14.3	14.3	16.9	13.5	12.1	16.4
TX	9.0	10.1	9.7	9.2	9.6	8.0	9.4
CA	7.4	7.6	7.3	11.6	10.7	7.7	10.0
DC	5.7	2.3	3.1	9.0	4.6	3.1	7.5
NJ	5.3	3.8	2.1	12.9	5.4	5.9	12.6
ND	-0.7	-0.6	-1.0	3.2	-3.1	1.9	5.2
VI	-4.0	-2.2	-3.5	-8.0	-3.9	-5.1	-6.7

Table 12 High-Low difference - Deviation from criterion method results in se(criterion) units

NO Scale:

	wwt	awt	aat	wwr	awr	aar
HI	0.15	0.15	-3.40	0.20	-0.05	-3.25
FL	-0.03	0.03	0.17	-0.17	-0.40	0.07
CA	-0.44	-0.07	-0.63	-0.11	-0.30	-0.70
NJ	-0.22	-0.14	0.44	0.11	-0.19	0.31
ND	-0.26	-0.71	-0.71	-0.45	0.21	-0.17
TX	0.09	0.16	-0.09	-0.50	0.03	-0.31
DC	-0.35	-0.19	0.97	-0.58	-0.81	0.58
VI	-0.32	0.09	0.76	-0.03	0.00	0.91

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
HI	0.48	-0.30	-5.19	-0.33	-0.41	-4.89
FL	0.17	0.17	0.91	-0.06	-0.46	0.77
TX	0.26	0.16	0.05	0.14	-0.23	0.09
CA	0.06	-0.03	1.27	1.00	0.09	0.79
DC	-0.97	-0.74	0.94	-0.31	-0.74	0.51
NJ	-0.31	-0.65	1.55	0.02	0.12	1.49
ND	0.01	-0.04	0.57	-0.35	0.38	0.87
VI	0.40	0.11	-0.89	0.02	-0.24	-0.60

Table 13 TV1 - TV6 difference deviated from national mean

ND Scale (National mean = 24):

	crit	wwt	awt	aat	wwr	awr	aar
NJ	8.9	9.2	9.1	11.2	9.9	9.4	10.3
ND	-3.3	-1.1	-2.6	-1.2	-2.5	-2.6	-2.3
CA	-3.6	-3.7	-2.7	-4.4	-4.1	-3.0	-4.0
FL	-4.5	-4.1	-4.5	-4.6	-4.5	-4.7	-4.3
TX	-9.4	-9.3	-9.0	-8.0	-9.5	-9.9	-8.5
HI	-9.4	-9.0	-9.8	-8.6	-9.8	-9.0	-9.6
DC	-17.7	-18.5	-16.0	-15.5	-17.6	-17.4	-13.9
VI	-27.3	-25.9	-22.8	-23.3	-26.0	-26.4	-22.7

DASP Scale (National mean = 24):

	crit	wwt	awt	aat	wwr	awr	aar
NJ	18.1	18.5	16.5	21.9	18.4	17.0	22.7
ND	2.6	4.4	4.0	7.2	4.6	2.8	4.9
CA	0.0	1.3	1.3	2.8	-0.7	1.5	2.5
TX	-0.5	-0.7	0.3	0.1	-3.7	-2.1	-1.6
HI	-3.1	-3.9	-4.1	-3.8	-5.4	-5.4	-4.5
FL	-3.6	-3.5	-5.4	0.5	-6.5	-3.2	-0.1
DC	-11.1	-8.6	-9.7	-8.2	-11.0	-8.3	-6.2
VI	-34.6	-34.5	-34.9	-20.8	-29.9	-32.1	-22.1

Table 14 TV1 - TV6 difference --- Deviation from criterion method results in se(criterion) units

NO Scale:

	wwt	awt	aat	wwr	awr	aar
NJ	0.08	0.05	0.59	0.26	0.13	0.36
ND	0.51	0.16	0.49	0.19	0.16	0.23
CA	-0.03	0.22	-0.20	-0.12	0.15	-0.10
FL	0.12	0.00	-0.03	0.00	-0.06	0.06
TX	0.03	0.11	0.39	-0.03	-0.14	0.25
HI	0.12	-0.12	0.24	-0.12	0.12	-0.06
DC	-0.13	0.27	0.35	0.02	0.05	0.60
VI	0.52	1.67	1.48	0.48	0.33	1.70

DASP Scale:

	wwt	awt	aat	wwr	awr	aar
NJ	0.09	-0.34	0.81	0.06	-0.23	0.98
ND	0.38	0.30	0.98	0.43	0.04	0.49
CA	0.28	0.28	0.60	-0.15	0.32	0.53
TX	-0.04	0.16	0.12	-0.65	-0.33	-0.22
HI	-0.19	-0.24	-0.17	-0.55	-0.55	-0.33
FL	0.02	-0.43	0.98	-0.69	0.10	0.83
DC	0.30	0.17	0.35	0.01	0.34	0.60
VI	0.03	-0.10	4.60	1.57	0.83	4.17

Appendix 1 Number of contrasts, number of 0 eigenvalues, and the number of components deleted due to computational instabilities

	# of contrasts	# of 0 eigenvalues	# of unstable PCs
CA	166	15	25
DC	166	20	27
FL	166	16	25
HI	166	19	24
NJ	166	16	28
ND	166	21	28
TX	166	27	15
VI	166	26	25

Appendix 2 Subgroup sample sizes for each jurisdiction

	CA	DC	FL	HI	NJ	ND	TX	VI
Males	1244	1003	1291	1341	1360	1279	1261	644
Females	1180	1132	1243	1210	1350	1206	1281	682

	CA	DC	FL	HI	NJ	ND	TX	VI
Whites	1091	54	1548	445	1789	2234	1175	20
Hispanics	818	192	398	264	363	70	926	265

	CA	DC	FL	HI	NJ	ND	TX	VI
High Abil	638	287	602	588	639	254	371	162
Low Abil	394	378	489	653	541	179	407	248

	CA	DC	FL	HI	NJ	ND	TX	VI
1-Hr TV	398	157	299	252	332	334	313	229
6-Hr TV	262	716	471	572	349	158	375	351