Article

Journal of Educational and Behavioral Statistics
2019, Vol. 44, No. 6, pp. 648–670
DOI: 10.3102/1076998619867105
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://jebs.aera.net

Research on Psychometric Modeling, Analysis, and Reporting of the National Assessment of Educational Progress

Andreas Oranje
Educational Testing Service

Andrew Kolstad
P20 Strategies, LLC

The design and psychometric methodology of the National Assessment of Educational Progress (NAEP) is constantly evolving to meet the changing interests and demands stemming from a rapidly shifting educational landscape. NAEP has been built on strong research foundations that include conducting extensive evaluations and comparisons before new approaches are adopted. During those evaluations, many lessons are learned and discoveries surface that do not often find their way into widely accessible outlets. This article discusses a number of those insights with the goal to provide an integrated and accessible perspective on the strengths and limitations of NAEP's psychometric methodology and statistical reporting practices. Drawing from a range of technical reports and memoranda, presentations, and published literature, the following topics are covered: calibration, estimation of proficiency, data reduction, standard error estimation, statistical inference, and standard setting.

Keywords: *achievement; assessment; item response theory; NAEP; psychometrics; survey research*

Introduction

Over the past three decades, the National Assessment of Educational Progress (NAEP) statistical methodology has been extensively studied and occasionally evolved within the core design principles that were introduced by Messick, Beaton, and Lord (1983). The basic design of group score assessments calls, within a particular domain (e.g., reading, mathematics, writing), for shorter collections of test questions at the individual student level in order to minimize participant burden. Statistical inferences are made, and results are reported at the group level and with respect to content domains that are much broader than what can be assessed at the individual level in the amount of available time. Therefore,

a matrix test design (e.g., Frey, Hartig, & Rupp, 2009) is employed that can meaningfully link many shorter collections of questions together across participating students by systematically overlapping some questions between collections and, as a result, between students. In addition, test takers, their teachers, and schools fill out survey questionnaires that solicit information about demographics and instructional practices. Responses to these questions and other collateral information about test takers and schools are used in the analysis and define many of the groups about which results are reported. Lastly, data analysis is carried out through psychometric models that allow for direct estimation of group-level test results without the need to produce scores for individual test takers. The models are discussed in greater detail below.

The goal of this article is to summarize and discuss the research on psychometric models, analysis techniques, and reporting over the past 30 years. This research has often been motivated by increases in the scope of the program including increasing the number of jurisdictions for which these group-level results are reported (i.e., wider state participation, urban district participation, and a Puerto Rico mathematics assessment), offering testing accommodations for students with disabilities and English language learners, adding content domains (e.g., technology and engineering literacy), and expanding on survey questionnaires. Subsequently, the reporting of assessment results moved from a contained and focused set of key results to the provision of user customizable web-based analysis tools to satisfy a wide range of interests. Yet, despite the fact many improvements have been implemented and many more are possible, the core methodology has proven to be robust throughout those 30 years.

Statistical Model for Proficiency Estimation

The NAEP analysis model (Mislevy, 1984, 1991; Mislevy, Johnson, & Muraki, 1992) is a latent regression of one or more (correlated) latent traits of interest onto student group indicators. For student i and a single, univariate, latent trait θ:

$$\theta_i = \gamma' x_i + \varepsilon_i, \tag{1}$$

where γ is a vector of regression weights, x_i a vector of student group indicators, and ε_i a normally distributed residual term. To report on student group means and distributional quantities, we are principally interested in the distribution of θ given student group indicators x_i and responses to test items y_i:

$$f(\theta|x_i, y_i) \sim P(y_i|\beta, \theta)\phi(\theta|\mu = \gamma' x_i, \sigma^2), \tag{2}$$

where P is a likelihood function containing the product of individual item probabilities and ϕ is a normal distribution function representing the population prior. NAEP makes use of standard item response theory (IRT) models with item parameters β, including the two- and three-parameter logistic (3PL) models for

dichotomous items (Lord & Novick, 1968) and the generalized partial credit model (GPCM) for polytomous items (Muraki, 1992).

As described elsewhere (von Davier, Sinharay, Oranje, & Beaton, 2007), the operational approach consists of carrying out item calibration and the estimation of scale score results for reporting groups in separate sequential steps. This separation to a large extent has been due to tractability (e.g., in terms of parameter identification) and computer processing feasibility, at the cost of some level of inconsistency in model assumptions between the two different analysis steps. The primary nature of research and, occasionally, change over the past two decades in statistical modeling and estimation has for the most part centered on the following two questions: (1) How to model multiple latent dimensions to characterize multiple subdomains within overall content domains as well as crosscutting concepts and practices (e.g., inquiry skills across the science content domain) where some items may load on multiple latent dimensions and (2) how to increase the number of "predictor" variables in the latent regression without over-fitting the data—in order to accommodate the increasing number of reporting groups of interest.

Item Calibration

As mentioned before, the likelihood function P contains the product of individual item probabilities, which are represented by logistic parametric functions under the IRT framework (Lord & Novick, 1968). For dichotomous constructed response items that are scored as either right ($Y = 1$) or wrong ($Y = 0$), a two-parameter logistic model is used:

$$P(Y = y|\theta) = \left(\frac{e^{Da(\theta-b)}}{1 + e^{Da(\theta-b)}}\right)^{y} \left(1 - \frac{e^{Da(\theta-b)}}{1 + e^{Da(\theta-b)}}\right)^{(1-y)}, \tag{3}$$

where D is a scaling constant that is used to better approximate a normal cumulative distribution with this logistic model, a is a discrimination parameter to signify how well an item differentiates between students of different abilities, and b is an indication of how difficult the item is. Both the a and b parameters are quantities that need to be estimated from the data in the first analysis step. For multiple-choice questions, a guessing parameter is added to this model (i.e., 3PL), and for polytomous items, location parameters are added (i.e., the GPCM).

The goal of estimating these parameters is to (1) place all the items on a common IRT scale that quantitatively represents performance levels with respect to a construct of interest and (2) make use of a scale that is consistent with the established trend scale in order to make comparisons over time. The assumption is made that if the same items function similarly across two adjacent assessments, then the scale that is defined by those items is the same across those adjacent assessments. The data from both assessments can be pooled, and a single set of item parameters can be estimated based on all the available data. In order to be

able to estimate those item parameters, an estimate for the mean and variance of θ is required, and the distribution cannot be assumed to be the same across the two pooled data sets since the population of test takers changes (e.g., fourth graders in 2015 are not the same as fourth graders in 2017). NAEP uses marginal maximum likelihood (e.g., Johnson, 2007) where θ is assumed to be a random effect that can be integrated over to obtain a marginal distribution.

Operational estimation of item parameters is conducted using a modified version 3 of the PARSCALE software program (Muraki & Bock, 1999). The integration is conducted using numerical quadrature, and the distribution is either assumed normal or approximated as a multinomial distribution on a fixed set of θ values (Muraki, 1992). This latter approximation was developed to allow for potentially nonnormal proficiency distributions. Following NAEP's operational procedure, first an estimation of item parameters is conducted that assumes a normal distribution shape for the proficiency distribution. Subsequently, using the item parameters from this solution as starting values, a new set of item parameter estimates is generated with the proficiency distribution modeled as a set of multinomial probabilities. The identifiability of the multinomial approach has been questioned as both person and item parameters are estimated concurrently (Aitkin & Aitkin, 2006a, 2006b). In response, Sgammato (2012) conducted a study showing that the concerns, while valid, may not generalize to the specific NAEP design and data and, therefore, did not warrant an overhaul of the analysis procedures. While operational procedures have remained the same, more flexible, parametric approaches to modeling the proficiency distribution during the item-parameter estimation phase is an area of continued interest (e.g., Xu, 2007).

In terms of new developments related to item calibration, significant attention has been paid to multidimensional models for proficiency. The National Assessment Governing Board (NAGB), a nonpartisan board that sets policy for NAEP, develops and publishes content frameworks for each subject of interest. These frameworks are documents that describe measurement objectives for each subject and form the blueprint, both in terms of content and format, for the assessment. Frameworks have traditionally specified measurement objectives by subdomain (e.g., literary and information subdomains within the reading overall domain), and scale scores are reported for each of those subdomains in addition to a weighted average of subdomains, usually referred to in NAEP as a composite scale, to represent the overall scale. The weights are prescribed by the frameworks.

Under the current approach, IRT calibrations are carried out separately for each of the subdomains. In other words, for each scale, a set of item parameters and a univariate distribution of the latent variable are estimated. One critique of modeling the latent variable distribution as a sequence of univariate distributions rather than as a multivariate distribution is that valuable information from correlations between subscales is ignored, thereby offering less efficient item parameter estimates than would be obtained with a fully multivariate approach. As mentioned in the introduction, crosscutting concepts and practices have been

introduced in frameworks more recently that are measured across subdomains (e.g., inquiry across science knowledge domains). This implies a shift from, in factor analytic terms, a strictly simple structure design to a more complicated multivariate latent structure where items may measure multiple latent variables. A well-known issue for multivariate models estimated with marginal maximum likelihood is that it typically requires exponentially increasing computer resources as the number of dimensions increases and the number of quadrature points increases. However, several efficient approaches have been explored that require less computation and that could be fruitful approaches for appropriately and efficiently modeling multivariate latent variables with items that contribute to the measurement of multiple dimensions of latent variables. Based on research and development by von Davier (2005a), Xu and von Davier (2006) applied the generalized diagnostic model to NAEP data and obtained good recovery of group score statistics (e.g., means and standard deviations were close and within random variation) relative to the regular operational procedures. This model uses latent classes (i.e., a discrete approximation to the proficiency distribution) as the underlying latent trait structure and is computationally less intensive than a continuous variable full information approach.

Another multivariate direction that has been pursued more recently is the application of bifactor models to NAEP data. Following Rijmen (2009), the bifactor model (Gibbons & Hedecker, 1992) has the distinct advantage that higher dimensional models can be fitted without the computational burden of high-dimensional integrations. This model and computational approach can be estimated with the multidimensional item response theory (MIRT; Haberman, 2013) program and was applied to data from the 2014 NAEP Technology and Engineering Literacy assessment. In this assessment, both domain-specific dimensions and practices are included, in addition to larger scenario-based tasks that might represent additional dimensions associated with the specifics of a task (i.e., task effects). The bifactor model was set up as a single overall primary factor and several secondary and tertiary factors that included subdomains, practices, or tasks. The results (Shu, Xu, & Jia, 2013) showed that task effects tend to significantly overwhelm the effects of either domain knowledge or practices in this case. This is an active area of continued research, and the model was not used operationally.

Group Score Estimation

In the second phase of the operational procedure, the item parameter estimates ($\hat{\beta}$) are treated as fixed and known. In Equation 2, we introduced the univariate latent regression probability function. For most NAEP subjects, a multivariate version is used:

$$f(\Theta|x_i, y_i) \sim P(y_i|\hat{\beta}, \Theta)\Phi(\Theta|\mu_i' = \Gamma x_i, \Sigma), \tag{4}$$

where Θ is the multivariate latent variable of interest and Φ is the multivariate population prior with mean vector μ_i' and covariance matrix Σ, which is common

across student groups. The second phase entails the estimation of group effects Γ and covariance matrix Σ in order to report on differences between groups of students. Similar to item calibration, estimation can be conducted by employing marginal maximum likelihood and the expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). During the maximization step, Γ and Σ are estimated following least squares equations from solving the full likelihood for Γ and Σ. For Γ this is

$$\hat{\Gamma} = (X'X)^{-1}X'\tilde{\Theta}. \tag{5}$$

Both this equation and the equation for Σ rely on two quantities, the provisional posterior mean vector and the provisional posterior covariance matrix, which have elements for each dimension k and k' that are estimated in the expectation step as follows:

$$\widetilde{\theta_{i,k}} = \int_{\theta_k=-\infty}^{\infty} \theta_k P_i(y_i|\hat{\beta},\Theta)\Phi(\Theta|\hat{\Gamma}x_i,\hat{\Sigma})d\theta, \tag{6}$$

$$\widetilde{\sigma_{i,kk'}} = \int_{\theta_k=-\infty}^{\infty}\int_{\theta_{k'}=-\infty}^{\infty} (\theta_k - \widetilde{\theta_{i,k}})(\theta_{k'} - \widetilde{\theta_{i,k'}})P(y_i|\hat{\beta},\Theta)\Phi(\Theta|\hat{\Gamma}x_i,\hat{\Sigma})d\theta. \tag{7}$$

Therefore, the main computational challenge is to evaluate the integrals in Equations 6 and 7. The GROUP program (Mislevy & Sheehan, 1987) was developed for univariate scales and used numerical quadrature to evaluate the integrals. For the multivariate case, a number of studies were conducted, and programs developed resulting in the currently operationally used program CGROUP (Thomas, 1993), which uses a LaPlace approximation to evaluate the multivariate integrals. The details of this and alternative approaches to the multivariate case are outside the scope of this article and well described elsewhere (e.g., von Davier, 2005a). More recent work has focused on answering two questions: (1) Can the multivariate integrations be performed more efficiently and accurately? (2) Can some of the model assumptions be relaxed or additional effects be included in the model to avoid making those model assumptions (e.g., including effects due to the hierarchical nature of the sample where students are inside schools, schools inside districts)?

To answer the first question, Cohen and Jiang (1999) developed a software program called AM. Among other capabilities, this program estimates the same multivariate latent variable model based on marginal maximum likelihood. However, instead of evaluating multivariate integrals, the group effect and variance parameters associated with each of the scales are estimated separately first and then correlations between the scales are estimated last. This approach is based on an analogy to a seemingly unrelated regressions (SUR) model with fully observed dependent variables (Zellner, 1962). von Davier (2005b) implemented

the SUR approach in the standard NAEP operational software in order to conduct a direct comparison. It was shown in a simulation that the full multivariate approach recovered simulated parameters often equally and sometimes significantly better than the SUR approach. The SUR approach was, therefore, not implemented operationally.

Based on Mislevy's (1984) work, Antal and Oranje (2007) investigated whether the use of fewer, carefully selected quadrature points during the expectation step would make models with more latent traits computationally feasible. A software routine using Gauss–Hermite quadrature (as opposed to rectangular quadrature) was developed and implemented in the operational software. A simulation study (Kong, 2012) showed that once the number of points was reduced enough to make it competitive with Thomas's LaPlace approximation in terms of computation time, parameter estimates were considerably less accurate than what could be obtained with the LaPlace approximation. Pursuing a similar goal of more efficient multivariate integrations, von Davier and Sinharay (2004) developed a stochastic EM algorithm based on importance sampling of the posterior distribution to estimate provisional posterior means and covariances. They found that in the aggregate, this method provides a viable alternative to existing operational methods, but that further research was needed before operational use to investigate some convergence issues and outliers in the conditional posterior distribution and to reduce computational resources required.

The second question about relaxing model assumptions has led to several methodological developments and variations that can be classified as (a) those that pursued estimating all parameters (i.e., both item parameters and the parameters of the latent regression) concurrently (and, therefore, not assuming item parameters to be known and fixed in a separate estimation stage with separate assumptions about the distribution of the latent variables) and (b) those that pursued introducing additional parameters in the latent regression model to model additional effects (and, therefore, no longer assuming those effects to be ignorable). ACER's ConQuest (Adams, Wu, & Wilson, 2015) allows for the joint estimation of item and population parameters by placing strong constraints on the measurement model in order to make the estimation problem tractable in a single step. Specifically, a Rasch model is used to model item functions and, subsequently, estimate group-level quantities of interest. In addition, von Davier and Sinharay (2009, 2010) estimated all model parameters concurrently with a Metropolis–Hastings algorithm showing that the method is able to recover parameters within random variation in a simulation and particularly with small item sets, but that some research still has to be done on monitoring convergence. Possibly one of the most complete proposals for a comprehensive model for NAEP that entails an integrated hierarchical measurement and population model is provided by Aitkin and Aitkin (2011, p. 53). They propose a model that has a provision for guessing (though generalized among items) random effects for schools and includes ethnicity as a model factor. They base their work on commercially available packages

and provide guidance on how to obtain correct standard errors. They apply their models to relatively older data that have modest missingness by design and show accurate estimates of student group means and standard errors. They also point out a number of smaller technical details beyond the scope of this article that would require further research as well as the need to apply these models to operational cases, which typically contain more items overall, fewer items per student, larger samples, and a different hierarchical structure.

Several studies have been conducted, which focus on introducing additional parameters to more faithfully represent the data. For example, in order to allow different student groups to have different variances, Thomas (2000) developed an alternative version of the operational program where instead of a single variance (i.e., σ^2), different variances for different (mutually exclusive) student groups could be estimated. Upon further inspection, this version had limited gain in terms of more accurately recovering population distributions unless all identified groups are significantly large. Thomas's work was taken further by von Davier and Yon (2004) who developed a generalized least squares estimator in which the homoscedasticity assumption was relaxed for the residual variances at the student level. They did not find a substantial gain for balanced designs, where every booklet yields approximately similar levels of reliability.

Another set of examples pertains to the introduction of additional parameters to model the hierarchical nature of the sample, where students are sampled from schools, schools are sampled from districts, and so on. Li, Oranje, and Jiang (2009) developed a hierarchical latent regression model, estimating random effects parameters to reflect the nested structure of the data. Comparison of a hierarchical model with the operational model in terms of means and standard deviations of student groups showed small differences (i.e., within statistical significance bounds). This finding reflects the fact that the sizable number of background variables in the NAEP operational model represent most of the within and between school effects that are explicitly parameterized in a hierarchical random effects model. Johnson (2002) and Johnson and Jenkins (2005) developed a fully Bayesian solution using Markov chain Monte Carlo methods to estimate a hierarchical version of the group score assessment model. They also estimated all parameters jointly by imposing relatively informative priors. Recovery of simulated effects was reasonable but not showing a significant improvement over the existing NAEP approach.

Student Group Variables and Data Reduction

Earlier, we introduced x as a vector of student group variables for which we want to report scale scores. These variables may come from school records (e.g., gender, race/ethnicity, disability, and English language learning status), student responses to questionnaires about a variety of topics such as access to resources and study habits, teacher responses to questionnaires about a variety of topics such

as teacher preparation and experience, and school administrator responses to questionnaires about a variety of topics such as governance and school culture. By including these variables in the latent regression analysis, the relationship between the groups of interest as defined by the x variables and proficiency is estimated. The estimated latent regression—specifically the model-implied posterior distribution of theta associated with each test taker—can be used to derive estimates of the proficiency distribution for the various reporting groups of interest. Subsequently, the resulting posterior distribution becomes the basis for reporting student group means. This posterior distribution can also be generated with few variables. However, as Mislevy (1984, 1985) points out, omitting indicators of student groups of interest from the set of x variables that define the latent regression model leads to potentially biased estimates of the proficiency distribution for those groups.

One goal of the NAEP operational approach is to include as many student groups as possible in a single latent regression model to create a single, canonical set of results that provides consistency for all data users. This goal cannot be fully achieved for several reasons. A potential issue common to all regression is multicollinearity and near-zero variances of independent variables. Early on, this was addressed manually by multiple sequential analyses and removing independent variables as necessary. However, the number of student groups of interest has grown significantly over the years with the introduction of additional or longer questionnaires in order to provide more context to the results. Subsequently, a manual process for resolving multicollinearity became impractical.

As a result, NAEP adopted the use of principal components to remove multicolinearity and near-zero variances of independent variables. NAEP's practice is to convert ordinal and nominal responses from the NAEP survey questionnaires and other test taker covariates into a series of dummy-coded contrasts. Covariates that can be treated as continuous are left in their original form. Principal components of these survey responses and covariates are then obtained based on a correlation matrix, effectively standardizing all variables to have unit variance. More precisely, the current approach first makes use of the SWEEP algorithm (Goodnight, 1978) to remove most multicollinearity and very small variances, followed by a principal component analysis using the smallest set of variables that represents 90% of the variance for the latent regression analysis. This process is automated and can be applied to national samples as well as for each state consecutively in state-level assessments.

Cohen and Jiang (2002) proposed that the group score estimation should be conducted one or a few variables at a time to ensure that inferences would never be made about variables not explicitly included in the model and, therefore, would not incur any secondary biases. There are some practical challenges and statistical issues associated with that proposed approach. As discussed (Mazzeo, Donoghue, Li, & Johnson, 2006) and shown with real and simulated data (Moran, Dresher, & Davis, 2007), one issue with this approach is marginal inconsistency. That is, if the interaction of two variables, A and B, is modeled in one estimation

and the results are marginalized over B (e.g., you obtain the results for A), these results are not guaranteed to be identical to a model where only the variable A is modeled. In addition, Cohen and Jiang's approach makes the fairly strong assumption that proficiency is normally distributed in all reporting groups of interest. Moran et al. show that this procedure cannot recover group score statistics accurately when this assumption is violated. When the cognitive items matrix design is relatively sparse, adding more student group indicators into the model beyond those needed to define the specific reporting group(s) of interest can improve the accuracy of the standard errors of the group score statistics for groups that were already included (Thomas, 2002).

Various alternative approaches to reducing the number of predictor variables in the latent regression model have been studied. For example, Oranje and Ye (2014) compared the aforementioned operational approach with various other approaches. One approach was to include key reporting variables directly (i.e., as a set of dummy codes) into the set of independent variables while using a residual principal component analysis on the rest. The goal was to make sure that the most important and frequently reported variables would be fully represented in the model rather than through a combination of principal components. A second approach was to use covariances rather than correlations to base the principal component analysis on. The goal was to avoid making the assumption that all independent variables have the same standard deviation. They also manipulated the percentage of variance retained in the principal component portion to investigate whether a smaller model could still represent the variance in all the variables sufficiently. In terms of the bias–variance trade-off, it was found that even the largest models in this study still showed the most favorable result, meaning that bias gains still outweighed variance inflation in very large models with many parameters. The covariance-based approach was not appreciably different from the correlation-based approach. Lastly and not surprisingly, including indicators of key student groups directly into the independent variable set lowered the bias for those variables, but at the cost of increased bias of the other variables. Finally, Johnson (2011) studied the use of the least absolute shrinkage and selection operator regression with item responses included in the model as an alternative for principal components and found that these methods yielded comparable results in terms of means and variances relative to standard errors for the situations typically encountered in operational NAEP.

Plausible Values

Based on the ideas of Little and Rubin (1987), plausible values are the primary conduit for calculating and reporting official NAEP results. They are multiple imputations generated based on the estimated latent regression model and provide a stochastic approximation to the group score results that are implied by the model and, in principle, directly calculable from that model. Moreover, they provide a convenient mechanism for estimating one component of the error

variance in NAEP reported statistics associated with the fact that the dependent variable of interest (i.e., NAEP proficiency) is not directly observable. Plausible values are not the scores of individual test takers and should not be interpreted as such. From an operational perspective, having a single set of plausible values that can be used to calculate the results for many different groups is convenient and assures internal consistency across the full set of student groups for which results are sought.

The aforementioned component of error variance (*B)* represented by the variability in plausible values is calculated as follows:

$$B(z_g^m) = \frac{1}{R-1}\sum_r (z_g^{m,r} - z_g^m)^2, \tag{8}$$

where R is the total number of plausible values, z_g^m is a statistic m (e.g., the mean, standard deviation, a percentile) of interest for group g and averaged over plausible values after calculating the statistic $z_g^{m,r}$ for each of the plausible values R. The most significant change to this methodology has been made recently increasing the number of plausible values (R) from 5 to 20 in order to improve the estimation of B. While it can be shown that five values lead to robust results for B for most student groups, the additional draws do provide increased estimation accuracy of the statistics of interest (Oranje & Freund, 2013) in smaller groups and states, particularly those who are more at the extremes of the distribution, where measurement variances (i.e., B) are relatively large.

Statistical Inference and Reporting

The NAEP program reports four main statistics for student groups: means, standard deviations, percentiles (10th, 25th, median, 75th, and 90th), and achievement-level percentages. Achievement levels are expert panel-determined cut points on the proficiency score scale, which signify the boundaries of intervals that are associated with basic, proficient, and advanced performance. Statistical inference is conducted primarily through pairwise t tests of student groups (e.g., male compared to female students) within a jurisdiction (e.g., nation, state, urban district) between 2 years or between two student groups or jurisdictions within a year (e.g., North Carolina compared to South Carolina in 2009). Changes to that process over the past two decades fall into four categories: (1) a more comprehensive reflection of what sources of error contribute to uncertainty, (2) significance testing, (3) flagging of results that may be less reliable, and (4) multiple pairwise comparison procedures. We will first review NAEP's operational standard error calculation.

In the previous section, the variance term B was introduced as the between-imputation variance intended to represent measurement variance. A second variance term S is used to represent the variance due to sampling. NAEP uses a multistage probability sampling design to sample students at random within schools, schools proportionally to their size (i.e., number of test takers eligible

for inclusion in the sample) within primary sampling units (PSUs), and PSUs proportionally to size. As a result, the sample is clustered, and design effects for major reported student groups typically are between 2 and 3. The combined standard error for a proficiency statistic z^m of interest for group g is the square root of V, which is calculated as follows (Mislevy et al., 1992):

$$V(z_g^m) = S(z_g^m) + \frac{R+1}{R} B(z_g^m), \tag{9}$$

for R total plausible values.

Following Hansen and Tepping's (1985) and Kovar's (1985) evaluation of various variance estimation methods, including Taylor Series expansion, bootstrapping, balanced repeated repetitions, and jackknifing, the calculation of S was determined to be most accurate under the jackknife repeated replications (JRR) method. Alternatives have been studied over the years. For example, von Davier (2005b) used White's robust estimator to obtain standard errors for the coefficients of the latent regression model with mixed success. In addition, Cohen and Jiang (2002) adapted Binder's estimator based on a Taylor series specifically for their one-variable-at-a-time methodology. Li and Oranje (2007) evaluated the Binder's method against the JRR and found that this method is only accurate for very small models and is inaccurate for larger models.

Sources of Error

In addition to sampling and measurement variance, some other sources of variance have received consideration over the past two decades, including item sampling. The idea of item sampling as a source of error is that there is a universe of content from which each instance of the assessment samples a set of items and tasks. This sample of the universe may be a significant source of error that should be included in standard error calculations. This is particularly true for shorter subdomain scales (e.g., the subdomain of geometry within the larger domain of mathematics) that are represented in the assessment by a relatively small set of items. Jiang, Cohen, Hsu, and Johnson (2005) studied the relative magnitude of this source of error and recommended the use of a double-jackknife approach developed by Cohen, Johnson, and Angeles (2000). This method successively leaves out both a student and an item unit, calculates statistics of interest (e.g., a mean for a particular student group), and uses the variability among these successive calculations as standard error due to both item and student sampling. This research was subsequently reviewed by an expert panel, and the review was summarized by Aitkin (2006). Aitkin concluded that the original formulation was technically problematic but that the notion of a double jackknife was not unreasonable and that further research was recommended to develop a sound and tractable method for estimating this source of variation and to include it in the standard error calculation.

Statistical Testing

As mentioned above, NAEP makes use of pairwise *t* tests to determine the statistical significance of differences between student groups, years, and jurisdictions. Up to 2005, NAEP treated all comparisons as though they were comparisons between independent samples. While for some of these comparisons (e.g., states to the nation, mutually exclusive groups among each other), such assumptions were not strictly correct because too large a variance is assumed, the program adopted this simple convention as a conservative strategy. It was assumed that standard errors under an independence assumption would be larger than under a dependence assumption (i.e., there is a positive correlation between dependent samples) and that it was safer to underreport on statistically significant results than overreport.

However, changes in the sampling design made the independence assumption less tenable as some dependencies could be rather substantial. As a result, adjustments to the way statistical tests are conducted were implemented for those cases where the samples of comparison were not strictly independent. Two big design changes were of particular impact: (a) combined samples and (b) urban districts. Prior to 2002, the state public school assessment program and the national assessment were conducted based on mutually exclusive samples. When No Child Left Behind was enacted, and states would risk losing federal funds if they would not participate in certain parts of NAEP (i.e., Grades 4 and 8 reading and mathematics), it became logical to integrate the two samples. From then on, the national *combined* sample was comprised of all the state public school samples plus a national private school sample. As a result, an inherent sample dependency was created when comparing a state to the nation, of which that state was also part. Starting in 2005, select large urban districts were representatively sampled and separately reported on. This created another level of dependency in comparisons between, for example, Chicago and Illinois, as Chicago is part of the Illinois sample. As a result, dependent *t* tests have been introduced. Note that in the combined national sample, states and large urban districts are weighted according to student population size.

Qualifying Results

Since the inception of modern-day NAEP (i.e., 1990s and on), several conventions have been adopted with the intention to qualify statistical testing results and inform the user of those results about the technical limitations that need to be taken into account. For example, two sample size–based conventions are the rules of 62 and 5. The rule of 62 was derived by adopting a convention that the minimum student sample size would have 0.80 power of detecting a 0.5 effect size difference in means controlling for Type I error at 0.05 and, for approximation purposes, assuming a design effect of 2 compared to simple random sampling. The rule of 5 determines the minimum number of PSUs or schools

(depending on the type of sample) required for reporting. The intention of this rule is to assure that the jackknife standard errors of the statistic of interest exceed a threshold for precision, particularly when the sample on which the results are based is clustered in a small number of sampling units (Johnson & Rust, 1993).

As mentioned earlier, one statistic of interest is the percentage of students at (or above) defined achievement levels on the score scale. In addition, the program reports on the percentage of students in various student groups defined by either school records (e.g., race/ethnic groups) or self-reports to questionnaires. Percentage metrics are typically noncontinuous (i.e., the denominator is a count) and bounded by 0 and 100. Therefore, computing a standard error following the aforementioned imputation and jackknife-based methods becomes problematic because those methods assume continuous, unbounded metrics and random error that is symmetric around a point estimate. For percentages in the middle of the range (e.g., around 50%), this may not be an issue. However, toward the extremes, this leads to cases where a symmetric confidence interval could surpass the bounds of the scale. NAEP has adopted a convention (National Center for Education Statistics, 2005) to suppress results for which this is the case. This includes comparisons between percentages if this is the case for at least one of the estimates in the comparison. Oranje (2006) studied various alternative approaches with asymmetric confidence intervals for percentage estimates and found that the Wilson (1927) interval is the most accurate in recovering appropriately sized intervals based on simulation. The program has started including (asymmetric) confidence intervals in an online data analysis tool, and comparisons of differences of percentages are based on the standard error of the difference rather than pooling two error terms and suppressing results.

Multiple Comparisons

Across many student groups (e.g., gender, race/ethnicity, type of location, answers to questionnaires), assessment years, and jurisdictions (i.e., states and urban districts), the number of possible comparisons can be sizable. Statistical testing is a critical component to determine which results can be reported as different. For every comparison that appears separately in a report, a Type I error rate of 5% is used. When many multiple comparisons are conducted simultaneously, the Type I error rate quickly approaches 1 across comparisons, and some correction is required. Historically, a Bonferroni correction has been applied where the Type I error rate is set at a much lower rate commensurate with the number of comparisons to account for capitalization on chance. Since then, NAEP has adopted an alternative methodology that controls the false discovery rate (FDR, Benjamini & Hochberg, 1995) rather than the cumulative probability of making Type I error. The main advantage is that the FDR method is statistically more powerful. As with most adjustment methods, the number of comparisons or *family size* needs to be determined. This is challenging in the sense that

determining the statistical significance of a comparison between two states or 2 years may depend on how many other states or years are included in the pairwise comparison and, therefore, how large the family size is. As a result, the program has adopted two conventions that limit the family size for those instances at the risk of occasionally overreporting of statistical significance: State to nation comparisons always carry a family size of 1 as do year-to-year comparisons.

Achievement Levels

The introduction of IRT models to NAEP in 1983 provided a way to connect items to NAEP scales in the same way that populations of students can be given scores. This opened the technological possibility of linking expectations for student performance in NAEP to ranges on the NAEP scales. Under the authority of the legislation that established NAGB in 1988, NAGB took on this responsibility and carried out an achievement level–setting process in each subject at the beginning of each trend line. In the early 1990s, the initial process of setting achievement levels on NAEP mathematics and reading scales was controversial. A nontechnical summary of this controversy can be found in Vinovskis (1998). In response to the controversy, Congress in 1994 changed NAEP's legislation to mandate periodic independent evaluations of NAEP and its achievement levels. Under this legislation, NAEP reports must make clear the "developmental" status of the achievement levels (later changed to "trial" status) until that status is ended by the Commissioner of Education Statistics who is to make a determination, on the basis of a mandated independent evaluation, that the achievement levels are "reasonable, valid, and informative to the public." While some of the issues in the controversies were political, the ones relevant in this context were technical.

The basis for performance expectations lies in collective judgments by subject matter experts who rely on policy definitions from NAGB of *Basic*, *Proficient* (defined in part as "competency over challenging subject matter"), and *Advanced* performance, as well as initial achievement-level descriptions based on the frameworks that define subject matter content coverage for each assessment. These experts are carefully selected and trained to make item-specific judgments of desired levels of performance. The initial achievement-level standards for NAEP's reading and mathematics assessments were based on two rating methods. For multiple-choice and short-answer questions, the judges estimated the proportion of Basic, Proficient, or Advanced students at the lower borderline of the category (as defined by the framework-based achievement-level descriptions) who would answer each item correctly. The cut score for these items was derived by averaging the item percentages and locating the score value along the IRT test characteristic curve that corresponds to that average percent correct. For polytomous items, the judges were asked to select from a set of sample papers (without knowledge of how the papers were scored) one that best represented the lower border of the achievement-level category. The item ratings of the selected

papers were averaged across the group of judges, and the cut scores for polytomous items were again located using the test characteristic curve. The two methods resulted in different cut scores; the differences were reconciled via an information-weighted average of the two methods. Details can be found in Bourque (1994), American College Testing (1993), and Luecht (1993) and the subsequent technical corrections in Loomis, Bay, and Chen (1996).

The principal technical problem with respect to the judgments of technical experts is that their understanding of the difficulty of the items may not correspond well with the empirically derived quantitative measures of item difficulty produced by IRT models. Subsequent to setting standards in reading and mathematics, and throughout the 1990s, NAGB's standard-setting methodologies in U.S. history, civics, and geography were evolving but were fundamentally derived from the same Angoff approach. Loomis and Bourque (2001) and Reckase (2000, 2001) provide overviews of this evolution. Since 2000, the achievement-level standards in Grade 12 mathematics, economics, technology and engineering literacy, and writing were set with the bookmark method (Kingston, Kahl, Sweeney, & Bay, 2001), at least in part because the bookmark method did not require as much time to conduct. With the bookmark method, the test items are arranged in order of difficulty in books given to the judges. The books contain all items (or a substantial proportion of them, as when different groups of judges are given different subsets of the item pool). The task then becomes placing a bookmark (i.e., finding the dividing line) between the items that students should be expected to answer correctly at the lower borderline of the Basic, Proficient, and Advanced levels and those they are not expected to be able to handle. The bookmarks can be associated with an IRT score through the response probability convention. While there is an element of arbitrariness about the choice of this convention (Loomis & Bourque, 2001; Kolstad, 1999), the value of a 65% chance of correctly answering a question was discussed by Zwick, Senturk, Wang, and Loomis (2001) and has become a de facto standard throughout the testing industry. In this approach, the various levels of partial credit are treated as separate items and included in the ordered item books. As a result, with the bookmark method, there is no issue of combining cut scores derived from different methods.

Documenting the conduct of the achievement-level setting process has been NAGB's responsibility, with little online technical information in NAEP's technical documentation on the Web. Technical reports explaining the process for each subject since 1996 are available on the NAGB website, and photocopies of the earliest 1992 and 1994 reports are available upon request to NAGB. During the 1990s, NAGB provided summaries of the standard setting process for inclusion as appendices in NAEP's printed technical reports.

The success of setting achievement levels on NAEP scales has been criticized, especially the earliest ones in reading and mathematics (National Academy of Education [NAE], 1993a, 1993b, 1996). Congressionally mandated independent

evaluations of NAEP (NAE 1997; National Research Council 1999) resulted in negative evaluations of the NAGB's achievement-level setting as a flawed process that did not encourage ending the trial status of the achievement levels, although the evaluations did encourage reporting trends against the unchanging metric of the achievement-level cut scores. A 2009 evaluation by the NAEP Evaluation Technical Working Group, published in a special issue of *Applied Measurement in Education* (volume 22, no 4), found that many users find reporting by achievement levels useful, that the procedures used are consistent with professional standards, but that additional external validity evidence is needed. The 2016 independent evaluation by the Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading recommended that only "once satisfactory alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores in NAEP mathematics and reading has been demonstrated, their designation as trial should be discontinued." The Commissioner responded to this evaluation by determining that the trial status of the achievement levels in reading and mathematics should be continued. At this time, all reports from NAEP are still legislatively required to include cautionary language about the trial status of NAEP's achievement levels.

Summary and Conclusion

In this article, we have provided an overview of the NAEP methodology and some of the psychometric and statistical methodology issues that have been studied over the past three decades. Occasionally, some changes to the NAEP operational methodology have been implemented alongside larger design changes of the program. More often, the existing operational methodology appeared very robust relative to potential alternatives.

We started with an overview of the operational item calibration methodology and discussed what forays into full MIRT have been made. As the interest in skills that cut across multiple knowledge domains increases, we expect that MIRT models will become important components of the methodology. The main questions for estimating more complex models will concern efficient estimation of a considerable number of parameters and how to connect that with the latent regression of student groups on proficiency. Subsequently, we discussed advances in the latent regression methodology itself, including different paradigms to approach the task and more minute changes to how the various quantities of interest are estimated by either relaxing or adding assumptions or trying out different estimation and approximation methods. The size of the model (e.g., number of student groups of interest) coupled with the desire to estimate and report on a single canonical and reproducible set of results means that the estimation task is sizable and grows with the expanse of student group variables. This area will need to focus and grapple with balancing model size with the ability to efficiently produce consistent results.

Underneath the core estimation methodology resides a distinct philosophy about linking items and trend assessments, which is based on common items between successive administrations. This philosophy is routinely challenged by programmatic and societal changes. For example, technology is already significantly influencing teaching, learning, and assessment. An important question for the program will be how to maintain a meaningful trend if the circumstances under which the assessment is assessed in terms of commonly shared technology are so starkly different from one point to the next. One way that such changes can be quantified, if still considered meaningful from a substantive perspective, is through appropriately accounting for additional sources of bias and variance. For example and as mentioned earlier, further research on the effect of item sampling across assessments might be worthwhile as well as linking error associated with changes of assessment mode. Another source of error variance not currently accounted for is the variability in human ratings of performance on constructed response items.

Finally, we discussed research on and some changes that have been made to the statistical inference and standard setting methodology. The program's transformation from providing a restricted set of precalculated results to an environment in which the public can obtain results about many student groups on-the-fly has also transformed the statistical testing practice. An important focus for the program might be to develop new paradigms for multiple comparisons that are consistent across different comparisons. This is especially important as the amount of data and possible comparisons will only increase with the advent of (a) digital assessment environments that can record all student's interactions with the test and (b) more elaborate, scenario-based tasks that invite and require a lot more interaction.

The adoption of IRT methods in the early 1980s was followed 10 years later by the development of methods to map expectations for student performance onto the NAEP scales in the form of achievement levels in reading and mathematics. These methods evolved and changed as expectations for student performance in U.S. history, civics, geography, science, writing, economics, and technology and engineering literacy were mapped to their corresponding NAEP scales. The program has made some important upgrades over the years and will need to continue to do so as different item and task types require different processes.

Authors' Note

The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Andreas Oranje prepared the work as employee of Educational Testing Service.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has been funded with Federal funds from the U.S. Department of Education under Contract number ED-IES-13-C-007.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised item response modelling software (Version 4) [Computer software]. Camberwell, Victoria: Australian Council for Educational Research.

Aitkin, M. (2006). *Review of enhancing the two-dimensional jackknife with application to the national assessment of educational progress and comments*. Prepared for the U.S. Department of Education, Washington, DC.

Aitkin, M., & Aitkin, I. (2006a, January). *Statistical issues in modern computing of IRT models*. Presentation to the NAEP Design and Analysis Committee, Princeton, NJ.

Aitkin, M., & Aitkin, I. (2006b, November). *Investigation of the identifiability of the 3PL model in the NAEP 1986 math survey*. Prepared for the U.S. Department of Education, Washington, DC.

Aitkin, M., & Aitkin, I. (2011). *Statistical modeling of the national assessment of educational progress*. New York, NY: Springer.

American College Testing. (1993). *Setting achievement levels on the 1992 NAEP in mathematics, reading, and writing: A technical report on reliability and validity*. Photocopy available upon request from Washington, DC: National Assessment Governing Board.

Antal, T., & Oranje, A. (2007). *Adaptive numerical quadrature for large scale assessments* (ETS Research Report No. RR-07-06). Princeton, NJ: ETS.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*, 289–300.

Bourque, M. L. (1994). The NAEP achievement level-setting process for the 1992 mathematics assessment Appendix G. In E. G. Johnson & J. E. Carlson (Eds.), *The NAEP 1992 technical report* (pp. 839–758). Washington, DC: National Center for Education Statistics. NCES 94–490. Downloaded from ERIC ED 376 191.

Cohen, J., & Jiang, T. (1999). Comparison of partially measured latent traits across nominal subgroups. *Journal of the American Statistical Association*, *94*, 1035–1044.

Cohen, J., & Jiang, T. (2002). *Direct estimation of statistics for the national assessment of educational progress (NAEP)*. Prepared for the U.S. Department of Education, Washington, DC.

Cohen, J., Johnson, E., & Angeles, J. (2000). *Variance estimation when sampling in two dimensions via the jackknife with application to the national assessment of education progress*. Washington, DC: American Institutes for Research.

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 (Methodological)*, *39*, 1–38.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues & Practice*, *28*, 39–53.

Gibbons, R. D., & Hedecker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Goodnight, J. H. (1978). *The SWEEP operator: Its importance in statistical computing* (SAS Technical Report No. R-106). Cary, NC: SAS Institute.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (ETS Research Report No. RR-13-32). Princeton, NJ: ETS.

Hansen, M. H., & Tepping, B. J. (1985). *Estimation of variance in NAEP* (Unpublished manuscript). Rockville, MD: Westat.

Jiang, T., Cohen, J., Hsu, Y.-C., & Johnson, E. (2005). *Enhancing the two-dimensional jackknife with application to the national assessment of educational progress*. Prepared for the U.S. Department of Education, Washington, DC.

Johnson, M. S. (2002). *A Bayesian hierarchical model for multi-dimensional performance assessments*. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, *20*. Retrieved from https://www.jstatsoft.org/article/view/v020i10/v20i10.pdf

Johnson, M. S. (2011). *Using item responses to select conditioning variables for NAEP*. Prepared for the U.S. Department of Education, Washington, DC.

Johnson, M. S., & Jenkins, F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the national assessment of educational progress* (ETS Research Report No. RR-04-38). Princeton, NJ: ETS.

Johnson, E. G., & Rust, K. F. (1993). *Effective degrees of freedom for variance estimation from a complex sample survey*. Paper Presented at the Annual Meeting of the American Statistical Association, Orlando, FL.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). The bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (Chapter 9, pp. 219–248). Mahwan, NJ: Lawrence Erlbaum.

Kolstad, A. (1999). *Standard-setting by the back door: "Mastery" as a criterion for mapping items onto IRT scales*. Paper Presented at the CCSSO National Conference on Large-Scale Assessment, Salt Lake City, UT.

Kong, N. (2012). *QPOST technical manual* (Unpublished ETS Technical Memorandum). Princeton, NJ: Educational Testing Service.

Kovar, J. (1985). *Variance estimation of non-linear statistics in stratified samples* (BSMD Working Paper No. 85-052E). Ottawa, ON: Statistics Canada.

Li, D., & Oranje, A. (2007). *Estimation of standard errors of regression effects in latent regression models using binder's linearization* (ETS Research Report No. RR-07-09). Princeton, NJ: ETS.

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large scale assessments. *Journal of Educational and Behavioral Statistics*, *34*, 433–463.

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

Loomis, S. C., Bay, L., & Chen, W.-H. (1996). The information weighting error. Appendix K. In N. L. Allen, D. L. Kline, & C. A. Zelenak (Eds.), *The NAEP 1994 technical*

report (pp. 915–946). Washington, DC: National Center for Education Statistics. NCES 97–897. Downloaded from ERIC ED 404 377.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the national assessment of educational progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (Chapter 7, pp. 175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *A theory of test scores* (Psychometric Monograph 7). Richmond, VA: Psychometric Corporation.

Luecht, R. M. (1993). Using IRT to improve the standard setting process for dichotomous and polytomous items Appendix I. In American College Testing (Ed.), *Setting achievement levels on the 1992 NAEP in mathematics, reading, and writing: A technical report on reliability and validity* (pp. 1–34). Photocopy available upon request from Washington, DC: National Assessment Governing Board.

Mazzeo, J., Donoghue, J., Li, D., & Johnson, M. (2006). *Marginal estimation in NAEP: Current operational procedures and AM*. Prepared for the U.S. Department of Education, Washington, DC.

Messick, S., Beaton, A., & Lord, F. (1983). *National assessment of educational progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton, NJ: National Assessment of Educational Progress.

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., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 83-84 technical report* (Report No: 15-TR-20; pp. 293–360). Princeton, NJ: Educational Testing Service.

Moran, R., Dresher, A., & Davis, S. (2007). *Results from NAEP marginal estimation studies*. Prepared for the U.S. Department of Education, Washington, DC.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E., & Bock, R. D. (1999). *PARSCALE*. Chicago, IL: Scientific Software.

National Academy of Education. (1993a). *Setting performance standards for achievement: A report of the National Academy of Education Panel on the evaluations of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels*. Stanford, CA: National Academy of Education.

National Academy of Education. (1993b). *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the evaluations of the NAEP Trial State Assessment: Background studies*. Stanford, CA: National Academy of Education.

National Academy of Education. (1996). *"Reading achievement levels." In Quality and utility: The 1994 Trial State Assessment in reading. The fourth report of the National Academy of Education Panel on the evaluation of the NAEP Trial State Assessment*. Stanford, CA: National Academy of Education.

National Academy of Education. (1997). *Assessment in transition: Monitoring the Nation's Educational Progress*. Stanford, CA: National Academy of Education.

National Center for Education Statistics. (2005). *Limiting comparisons involving extreme percentages*. Retrieved February 28, 2017, from https://nces.ed.gov/nationsreportcard/tdw/analysis/infer_guidelines_extreme.aspx.

National Research Council. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Oranje, A. (2006). *Confidence intervals for proportion estimates in complex samples* (ETS Research Report No. RR-06-21). Princeton, NJ: ETS.

Oranje, A., & Freund, D. (2013). *20 Versus 5 Plausible Values*. Presented to the U.S. Department of Education, Princeton, NJ, ETS.

Oranje, A., & Ye, L. (2014). Population model size, bias, and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 203–228)*. Boca Raton, FL: Taylor and Francis.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts*. Iowa City, IA: ACT.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (Chapter 6, pp. 159–173). Mahwah, NJ: Lawrence Erlbaum Associates.

Rijmen, F. (2009). *Three multidimensional models for testlet based tests: Formal relationships and empirical comparisons* (ETS Research Report No. RR-09-37). Princeton, NJ: ETS.

Sgammato, A. S. (2012). *Alternative scaling options for NAEP*. Prepared for the U.S. Department of Education, Washington, DC.

Shu, Z., Xu, X., & Jia, Y. (2013). *The application of bi-factor model in NAEP operational settings*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factorized likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309–322.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, *25*, 351–372.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, *67*, 33–48.

Vinovskis, M. (1998). Developing NAEP performance standards. In *Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB)* (Chapter VIII, pp. 41–57, 84–89). Washington, DC: National Assessment Governing Board. Retrieved from https://www.nagb.org/content/nagb/assets/documents/publications/95222.pdf.

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *Recent additions to the MGROUP set of programs* (Research Memorandum No. RM-05-02). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2004). *Applications of the stochastic EM method to latent regression models* (Research Report No. RR-04-34). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2009). *Stochastic approximation methods for latent regression item response models* (Research Report No. RR-09-09). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*, 174–193.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). Statistical procedures used in the national assessment of educational progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. *26*, pp. 1039–1055). Amsterdam, the Netherlands: Elsevier.

von Davier, M., & Yon, H. (2004). *A conditioning model with relaxed assumptions*. Paper Presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209–212.

Xu, X. (2007). *Estimating latent ability distribution by using general skew elliptical distributions*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Xu, X., & Davier, M. von (2006). *Cognitive diagnosis for NAEP proficiency data* (Research Report No. RR-06-08). Princeton, NJ: ETS.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*, 348–368.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. (2001). An investigation of alternative methods for item mapping in the national assessment of educational progress. *Educational Measurement: Issues and Practice*, *20*, 15–25.

Authors

ANDREAS ORANJE is a general manager at Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08542; email: aoranje@ets.org. His research interests are psychometric design and statistical analysis of National Assessment of Educational Progress, game-based assessment, and applications of artificial intelligence in education.

ANDREW KOLSTAD is a principal statistician at P20 Strategies, LLC, 700 Woodside Parkway, Silver Spring, MD 20910; email: ajk95@columbia.edu. His research interests are statistical and psychometric analyses of National Assessment of Educational Progress and other assessment surveys.

Manuscript received August 9, 2017
First revision received February 14, 2018
Second revision received June 30, 2018
Accepted February 20, 2019