Research Report

Variance Estimation for NAEP Data Using a Resampling-Based Approach: An Application of Cognitive Diagnostic Models

Chueh-an Hsieh

Xueli Xu

Matthias von Davier

November 2010

ETS RR-10-26

Variance Estimation for NAEP Data Using a Resampling-Based Approach: An Application of Cognitive Diagnostic Models

Chueh-an Hsieh
Michigan State University

Xueli Xu and Matthias von Davier
ETS, Princeton, New Jersey

November 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel Eignor

Technical Reviewers: Yue Jia and Jiahe Qian

Abstract

This paper presents an application of a jackknifing approach to variance estimation of ability inferences for groups of students, using a multidimensional discrete model for item response data. The data utilized to demonstrate the approach come from the National Assessment of Educational Progress (NAEP). In contrast to the operational approach used in NAEP, where plausible values are used to make ability inferences, the approach presented in this paper reestimates all parameters of the model, and makes ability inferences based on replicate samples of the jackknife without using plausible values.

Results of the standard errors are presented for estimates of group means, total means, and other statistics used in official reporting by NAEP. Differences in results between this approach and the operational approach are discussed.

Key words: NAEP, jackknife, general diagnostic model, variance estimation, plausible values

Acknowledgments

The authors would like to thank Dan Eignor, Yue Jia, Frank Rijmen, and Andreas Oranje for their comments and suggestions, and Ruth Greenwood and Kim Fryer for their assistance in copyediting. The authors would also like to thank Steve Isham for providing the data used in this study.

Table of Contents

List of Tables

Page

1. Background

1.1 National Assessment of Educational Progress (NAEP)

As an ongoing national research study, the National Assessment of Educational Progress (NAEP) is designed to provide national and state information on the academic performance of America's students (fourth, eighth, and twelfth graders) in various subjects, such as reading, mathematics, writing, science, and other subject areas. Often referred to as "the nation's report card," NAEP is administered by the U.S. Department of Education and the National Center for Education Statistics (NCES). It includes a range of surveys and assessments, which provide information on students' educational experiences, teachers' characteristics and practices, and school climate.

Like many national surveys, NAEP has adopted a complex sampling design to select student participants in the assessments. The major feature of the complex sample design includes cluster sampling (utilizing the differential sample selection characteristics) and sampling weights (including adjustments for school and student nonresponse and poststratification). As a major source of uncertainty, sampling variability provides information about how much variation in a given statistic would likely occur if another equivalent sample of individuals were observed (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001). Another important source of variability of NAEP scores is measurement error. Since items in NAEP assessments are administered according to a partially balanced incomplete block (pBIB) design, each student responds to only relatively few items. Thus, the uncertainty in estimation of proficiency is also a variability component due to the imprecision in the measurement of the scale scores (Johnson, 1989; Li & Oranje, 2007; Mazzeo, Donoghue, Li, & Johnson, 2006; Qian et al., 2001).

1.2. Objective of the Inquiry

A major goal of NAEP is to provide various ability inferences of the target population as well as the subpopulations of American youth. Since 1984, NAEP has reported these academic results using item response theory (IRT) models (Lord & Novick, 1968; Rasch, 1960) and latent regression models (Mislevy, 1991). The IRT models are used to calibrate the cognitive items, and the latent regression models are used to make inferences on the latent abilities.

Operationally, through the use of the software CGROUP,1 population-related ability estimates, such as subpopulation means, achievement levels, and score distributions for various reporting groups, are obtained from examinees' item response data and background data

(Mazzeo et al., 2006; Mislevy, 1991; von Davier, Sinharay, Oranje, & Beaton, 2007). This marginal estimation approach involves two stages: the parameters of a latent regression model are estimated in the first stage, assuming the item parameters are fixed; then this model, with its estimated parameters, is used to generate a set of plausible values (Mislevy, 1991). These plausible values are used to obtain the estimates for means, standard deviations, percentiles, and other summary ability inferences. A jackknifing approach coupled with the plausible values is adopted in NAEP operation to obtain estimation error for different population statistics (Johnson & Rost, 1992).

One consequence of ignoring the complex sample design is that the magnitude of the standard error of group-level statistics tends to be underestimated. It has been argued that the effect of ignoring the complex structure on the parameters of interest is relatively large in an NAEP operationally saturated model. In some situations, the effect is even more substantial (Mazzeo et al., 2006, pp. 68–69). This finding may be the result of assuming common-variance across subpopulations embedded in the latent regression models, and this effect may be alleviated by using a model that allows for the estimation of group-specific variances (Mazzeo et al., 2006; Thomas, 2000; von Davier, 2003). The general diagnostic model (GDM; von Davier, 2005) is such a model that allows for different ability variance assumptions for different subgroups (Xu & von Davier, 2006, 2008). In addition, the GDM makes it possible to estimate the item parameters and the parameters in the regression models simultaneously and repeatedly for jackknifing samples, due to the parsimonious feature of the model. In contrast, current NAEP operation does not have the capability to perform simultaneous estimation for jackknifing samples. Thus, the primary goal of this study is to use GDM, assuming a multiple-group population model, to obtain the estimation error based on a jackknife resampling procedure and compare it with the operational results.

2. General Diagnostic Models (GDM)

The GDM (von Davier, 2005) contains a large array of statistical models such as latent class analysis (LCA) models (Lazarsfeld & Henry, 1968; Goodman, 1974; McCutcheon, 1987) as well as discrete latent trait models with prespecified skill profiles and levels, and multidimensional IRT (MIRT) models (Ackerman, 1994, 1996). For instance, the GDM can be used to perform multiple classifications of examinees based on their response patterns with respect to skill attributes. Using ideas from IRT, log-linear models, and LCA, GDM can be

viewed as a general modeling framework for confirmatory multidimensional item response models (von Davier, 2005, 2007; von Davier & Rost, 2006; von Davier & Yamamoto, 2004). Within this comprehensive framework, many well-known models in measurement and educational testing—such as the unidimensional and multidimensional versions of the Rasch model (RM; Rasch, 1960), the two-parameter logistic item response theory (2PL-IRT) model (Lord & Novick, 1968), and the generalized partial credit model (GPCM; Muraki, 1992), together with a variety of skill profile models—are special cases of the GDM (von Davier, 2005).

In the following analyses, we will apply a compensatory GDM and use the software of *mdltm* (von Davier, 2005) to estimate the parameters of the model. In addition, we adopt a log-linear smoothing technique to facilitate the estimation of the latent skill space (Xu & von Davier, 2008). Through the use of a log-linear smoothing method, not only will the number of estimated parameters (associated with the latent skill distribution) be reduced substantially, but the interrelationship among distinct latent skills will also be well accounted for. In the software *mdltm*, the expectation-maximization algorithm (EM; Dempster, Laird, & Rubin, 1977) is implemented and used for parameter estimation. This implementation enables one to use standard tools from IRT for scale linking, deriving measures of model goodness of fit, assessing item and person fit, and estimating parameters (von Davier, 2005).

2.1 The Logistic Formulation of a Compensatory General Diagnostic Model (GDM)

In this section, we introduce the logistic formulation of the compensatory GDM applied in this study. The probability of obtaining a response x in the GDM is given as follows:

$$P\left(X_i = x \middle| \vec{\beta}_i, \vec{q}_i, \vec{\gamma}_i, \vec{a}, c\right) = \frac{exp\left[\beta_{xic} + \sum_{k=1}^{K} x\gamma_{ikc}\alpha_k\right]}{1 + \sum_{y=1}^{m_i} exp\left[\beta_{yic} + \sum_{k=1}^{K} y\gamma_{ikc}\alpha_k\right]}, \quad (1)$$

where x is the response category for each item i ($x \in \{1,2,...,m_i\}$); $\vec{a} = (a_1,...,a_K)$ represents a K-dimensional skill profile containing discrete, user-defined skill levels $a_k \in \{s_{k1},...,s_{kl},...,s_{kL_k}\}$, for $k = 1,...,K$; $q_{i\cdot} = (q_{i1},...,q_{iK})$ are the corresponding Q-matrix entries relating item i to skill k ($q_{ik} \in (0,1,2...)$, for $k = 1,...,K$); the parameters β_{xic} and $\gamma_{ikc} = (\gamma_{i1c},...,\gamma_{iKc})$ are real-valued

thresholds and *K*-dimensional slope parameters, respectively, and *c* is the group membership indicator. For model identification purposes, the researcher can impose necessary constraints on $\sum_k \gamma_{ikc}$ and $\sum \beta_{xic}$; also, with a non-zero *Q*-matrix entry, the slopes $\gamma_{ik\cdot}$ help determine how much a particular skill component in $\vec{a} = (a_1,...,a_K)$ contributes to the conditional response probabilities for item *i* given membership in group *c*. For multiple group models with a common scale across populations, the item parameters are constrained to be equal across groups, so that $\beta_{ixc} = \beta_{ixg} = \beta_{ix}$ for all items *i* and thresholds *x* as well as $\gamma_{ikc} = \gamma_{ikg} = \gamma_{ik}$ for all items *i* and skill dimensions *k*.

2.2 The Log-Linear Smoothing of Latent Class Space

In this section, we introduce the log-linear smoothing of the latent skill space predefined by the design matrix. Suppose we have *K* skills/attributes, the probability of a certain combination of these skills can be approximated by:

$$\log(P_g(\alpha_1,\alpha_2,...,\alpha_K)) = \mu + \sum_k \beta_{k,g}\alpha_k + \sum_k \gamma_k \alpha_k{}^2 + \sum_{i\neq j} \delta_{ij}\alpha_i\alpha_j, \qquad (2)$$

where μ , β_k , γ_k and δ_{ij} are parameters in this log-linear smoothing model, and g is a group index (Xu & von Davier, 2008; Haberman, von Davier, & Lee, 2008).

3. Sample and Data Sources

In order to obtain a representative sample, approximately 191,000 fourth graders from 7,600 schools were sampled and assessed in the NAEP 2003 reading assessment. Given the fulfillment of the minimum guidelines, results are presented for the nation, 50 states, and three jurisdictions that participated in the 2003 assessment, and for nine districts that participated in the Trial Urban District Assessment (TUDA; Donahue, Daane, & Jin, 2005). In addition, unlike the results obtained from participating states and other jurisdictions, the national results reflect both public and nonpublic school student performance. Generally, NAEP reports not only the overall results, but also the performance of various subgroups of students, where the statistics such as average scores and achievement-level percentages are the foci of interest.

Developed by the National Assessment Governing Board (NAGB), two reading contexts[2] and four reading aspects[3] were specified in the framework of 2003 reading assessment to

evaluate reading performance of fourth graders, such as population related means and standard deviation as well as percentiles. In order to minimize the burden on any individual student, NAEP uses matrix sampling, where each student is administered a small portion of the entire assessment. For instance, in fourth grade, students were given a test booklet that consists of two 25-minute blocks, and the item types included multiple-choice, short constructed-response, and extended constructed-response items. In addition, students were asked to complete two sections of background information questions (Donahue et al., 2005). The two reading contexts—reading for literary experience and reading to gain information—are currently taken as two subscales of psychometric analysis of NAEP Grade 4 assessments. These two subscales are denoted by skill 1 and skill 2 in this study, respectively.

Similar to 2003, in 2005, a nationally representative sample of more than 165,000 fourth-grade students participated in the assessment. The national results were based on a representative sample of students in both public and nonpublic schools.[4] The framework used for the NAEP 2005 reading assessment is the same as that used in 2003 (Perie, Grigg, & Donahue, 2005; see Table 1).

Table 1

The Number of Items in National Assessment of Educational Progress (NAEP) 2003 and 2005 Reading Assessments

Year	Subscales	Response categories in the item			Total
		Multiple-choice	Short constructed-response	Extended constructed-response	
2003	Reading for literary experience	40	8	3	51
	Reading to gain information	40	8	3	51
2005	Reading for literary experience	41	5	4	50
	Reading to gain information	35	11	3	49

4. Variance Estimation

In survey practice, the simple random sampling assumption is often violated. Thus, corrections to the calculation of desired statistics are needed. When the variance estimation of a nonlinear statistic becomes the focus of interest, this correction can be made in two plausible ways. One is the Taylor series linearization method: using a linear estimator to approximate the nonlinear one while accounting for complex sample design (Binder, 1983; Li & Oranje, 2007;

Williams, 2000). The other is the replication method: recomputing the statistic of interest using different, comparable sets to the original sample to measure the variance of the parameter estimator (Fay, 1989; Rust, 1985).

In the present study, we apply the resampling-based approach. Resampling techniques, such as the jackknife, balanced repeated replication (BRR), the methods of random groups, and the bootstrap were used in earlier developments in variance estimation (Efron, 1982; Rust, 1985; Rust & Rao, 1996; Johnson, 1989). By permitting fractional weighting of observations, the class of replication methods becomes considerably broader and more flexible (Fay, 1989). That is, through associating replicate weights with the characteristics of the observed sample cases, this replicate weighting approach lends itself particularly well to the analysis of data with highly complex design features (Dippo, Fay, & Morganstein, 1984).

The NAEP uses a modified BRR, derived from the jackknife procedure (Miller, 1974), to obtain the variance estimate of a statistic. The student replicate weights (SRWTs) in jackknife samples are derived based upon adjustments to the initial base weight. Examples of the adjustments may include nonresponse, trimming, poststratification, and the probability of selection for each primary sampling unit (Allen, Donoghue, & Schoeps, 2001). In NAEP 2003 and 2005 reading assessments, there were 62^5 jackknife samples with different sets of SRWTs. Consequently, the estimated variance of parameter estimate, *t*, was calculated by aggregating these 62 squared differences, $\hat{v}(t) = \sum_{t=1}^{62} (t_i - t)^2$, where t_i denotes the estimator of the parameter obtained from the *i*th jackknife sample (Qian et al., 2001). For further discussion of the variance estimation procedure used by NAEP, interested readers may refer to the paper by Johnson (1989, p. 315).

5. Empirical Evaluation

The GDM with both a single-group and a multiple-group assumption was applied to analyze the data. Under a single-group assumption, all students are assumed to belong to a single population with one latent skill distribution, while under a multiple-group assumption, different latent skill distributions are allowed for different groups. In this study, the multiple-group variable is defined by race/ethnicity. Four ethnicity groups are distinguished to form the different levels of this variable: White group, Black group, Hispanic group, and Asian/Pacific Islander group. As shown in Tables 2 and 5, compared to the results from the single-group assumption,

the results from using the multiple-group assumption show better fit in terms of several fit indexes, such as the Bayesian information criterion (BIC; Schwartz, 1978), Akaike information criterion (AIC; Akaike, 1974), and log-likelihood. Hence, in this study, the results from the multiple-group assumption, such as group means and standard deviation as well as the estimation error of the group mean, are compared with the results from NAEP operational analysis. The scale used in these comparisons is the one obtained from IRT calibrations, not the one converted to the NAEP reporting scale.

5.1 National Assessment of Educational Progress (NAEP) 2003 Reading Assessment for Fourth-Grade Students

The model fit indexes under different assumptions are shown in Table 2. One can observe that the GDM with a race-group assumption has better model fit. Thus the GDM coupled with the race-group assumption is used in comparison to the operational results.

Table 2

Model Evaluation Based on National Assessment of Educational Progress (NAEP) 2003 Reading Assessment

Model	# of parameters	-2 * log-likelihood	AIC per person	BIC
Single-group analysis	240	4,247,410	.607	4,250,328
Race-group analysis	960	4,215,853	.603	4,227,528

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion.

The mean and standard deviation for race/ethnicity subgroups are shown in Table 3. Similar patterns are shown in the results from using the GDM with race-group assumption and from NAEP operation. That is, from high to low score, the four racial groups have the following order: White group, Asian/Pacific Islander group, Hispanic group, and African American group. In addition, we can observe differences in the means for subgroups, and the differences are relatively large in subgroups of small sample size (such as in the Asian group). Moreover, the standard deviation estimates are smaller for White and Asian students when using the current approach, while on the contrary, the standard deviation estimates are larger for Black and Hispanic students when using the current approach.

Table 4 shows the comparison between the standard errors associated with the group mean estimates. The standard errors from using the current approach are slightly larger than those from the operation.

Table 3

The Mean and Standard Deviation for Ethnicity Subgroups in 2003 Assessment

	The results from using GDM				NAEP operational results			
	Literary subscale		Information subscale		Literary subscale		Information subscale	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
White	0.689	0.938	0.577	0.982	0.691	0.956	0.575	1.002
African American	-0.140	1.054	-0.351	1.070	-0.144	1.016	-0.349	1.031
Hispanic	-0.046	1.065	-0.290	1.099	-0.059	1.034	-0.282	1.051
Asian/Pacific Islander	0.571	0.991	0.495	0.987	0.613	1.030	0.478	1.083

Note. GDM = general diagnostic model, NAEP = National Assessment of Educational Progress.

Table 4

Standard Errors for Subgroup Mean Estimates in 2003 Assessment

White group (sample size 118,061)			
	Using GDM	NAEP operation	Ratio of GDM to the operation
Literary	.007	.006	1.167
Information	.008	.006	1.333
Black group (sample size 35,308)			
	Using GDM	NAEP operation	Ratio of GDM to the operation
Literary	.017	.011	1.545
Information	.018	.011	1.636
Hispanic group (sample size 23,839)			
	Using GDM	NAEP operation	Ratio of GDM to the operation
Literary	.021	.016	1.312
Information	.019	.017	1.118
Asian/Pacific Islander group (sample size 8,223)			
	Using GDM	NAEP operation	Ratio of GDM to the operation
Literary	.032	.033	0.970
Information	.038	.033	1.151

Note. GDM = general diagnostic model, NAEP = National Assessment of Educational Progress.

5.2 National Assessment of Educational Progress (NAEP) 2005 Reading Assessment for Fourth-Grade Students

The model fit indexes are shown in Table 5. The GDM with a race-group assumption has a better fit than a single-group assumption. Hence the race-group analysis is used in the comparison with the operational results.

Table 5

Model Evaluation Based on National Assessment of Educational Progress (NAEP) 2005 Reading Assessment

Model	# of parameters	-2*log-likelihood	AIC per person	BIC
Single-group analysis	235	3,650,627	.610	3,653,452
Race-group analysis	940	3,625,153	.606	3,636,450

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion.

The mean and standard deviation for racial subgroups are shown in Table 6. Large differences in ability estimates are evident with respect to the literary-experience subscale between the operation and current approaches. In addition, the standard deviations for White and Asian/Pacific Islander students tend to be smaller when using the current approach rather than the operational approach; while the standard deviations for Hispanic and Black students are larger compared to the operation.

Table 7 shows the comparison between the estimation errors for the group mean in 2005 data. Again, we observe that the standard errors for group means from the current approach are slightly larger than those from the NAEP operation.

6. Summary and Discussion

The application of the GDM in this paper is not focused on the detection of skills measured by the NAEP assessment, but on the improvement of ability estimates by borrowing information across subscales defined by the framework and on making ability inference based on a multiple-group population model. Compared with the NAEP operational analysis, where hundreds of background variables are used to extract group ability estimates, this approach is much more parsimonious.

Table 6

The Mean and Standard Deviations for Subgroups in 2005 Assessment

	The results from using GDM				NAPE operational analysis			
	Literary subscale		Information Subscale		Literary subscale		Information Subscale	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
White group	0.880	0.872	0.501	1.027	0.910	0.930	0.500	1.028
African American group	0.177	1.081	-0.422	1.067	0.119	0.970	-0.410	1.049
Hispanic group	0.264	1.087	-0.340	1.112	0.206	1.007	-0.341	1.102
Asian/Pacific Islander group	0.866	0.921	0.482	1.020	0.917	1.000.	0.462	1.114

Note. GDM = general diagnostic model, NAEP = National Assessment of Educational Progress.

Table 7

Standard Errors for Subgroup Means in 2005 Assessment

White group (sample size 99,425)			
	Using GDM	NAEP operation	Ratio of GDM to operation
Literary	0.007	0.005	1.400
Information	0.008	0.006	1.333
Black group (sample size 27,897)			
	Using GDM	NAEP operation	Ratio of GDM to operation
Literary	0.012	0.008	1.500
Information	0.014	0.009	1.555
Hispanic group (sample size 25,122)			
	Using GDM	NAEP operation	Ratio of GDM to operation
Literary	0.016	0.014	1.143
Information	0.016	0.015	1.067
Asian group (sample size 7,706)			
	Using GDM	NAEP operation	Ratio of GDM to operation
Literary	0.024	0.019	1.263
Information	0.026	0.022	1.182

Note. GDM = general diagnostic model, NAEP = National Assessment of Educational Progress.

The primary goal of this study was to obtain the estimation error of subgroup ability means and standard deviations under the GDM framework. Specifically, 62 jackknife samples coupled with different sets of weights, which are utilized in NAEP operation, are used in our current procedure. The results have shown that the estimation errors for racial subgroup means are slightly larger in the current approach than those in the operation. This increase may be due to the fact that the uncertainty in the item parameters is ignored in NAEP operation.

A number of differences were noticed between the approach taken using the GDM and the results using operational procedures. For one, the operational approach assumes normality in the conditional distribution of the latent trait, given the item responses and a large number of the background variables (von Davier, 2003). In contrast, the GDM approach does not assume a particular form of multidimensional ability distributions. As mentioned above, the GDM approach estimates item parameters of the MIRT model jointly and estimates the population model, a mixture of multiple (in this case, four known) populations concurrently with the item parameters. Most importantly, the item parameters in the operational analysis are assumed to be fixed and known, both for the purposes of estimating the population model and for the jackknife replications, while our proposed approach reestimates the item parameters and population distributions for each of the 62 jackknife samples. The capability of reestimating all parameters used in the GDM enables one to implement a complete jackknife procedure, which results in relatively large estimation errors for the group ability means.

The application of the GDM to the NAEP assessment data is not limited to what has been shown in this paper. In fact, the GDM is able to (a) facilitate the dimensionality exploration of the NAEP assessment (von Davier, 2005) and (b) reduce the number of background covariates when making inferences on group ability estimates. For example, the multiple-group assumption under the GDM allows for possibly different ability distributions (certainly different variance structures) across groups. This heterogeneity of variance structure may reduce the secondary bias in the group mean estimates (Thomas, 2000). This ongoing research study is geared toward expanding analysis and reporting alternatives for NAEP.

The complexity of the latent ability space introduces corresponding complexities into the statistical modeling and score reporting. In practice, because data-driven model specification is often messy, a high level of expert judgment is needed in formulating appropriate models. Moreover, these inferences must be communicated in ways that are of the most use to all

stakeholders. For instance, recently, Xu (2007) conducted an investigation to examine whether the monotonicity property can generally be sustained in GDM so that simple data summaries (e.g., the observed total score) can help inform the ordered categories of the latent trait and lead to the reporting of valid and reliable scaled scores. Obviously, in recent years, a parametrically complex IRT modeling framework has been called for. However, a general principle that the researcher should always keep in mind is the requirement that the assessment's purposes be fulfilled with the minimal degree of complexity. That is, in applying the principle of parsimony to the cognitive diagnosis model and skill assessment system, the model must be complex enough to provide sufficient skill information and still be parsimonious enough for the obtained skill information to meet user needs (DiBello & Stout, 2007; Haberman & von Davier, 2007).

Finally, the question of whether the larger variance estimates were observed because the GDM approach was carried out with a complete jackknife, that is, using recalibrations of item parameters and reestimated population models, rather than a jackknife based on imputations from a model that uses the complete sample, needs further investigation. If this was indeed the reason for observing larger variance estimates when using the GDM approach rather than the operational processes, an inquiry is needed into whether the added portion of variance reflects the true sampling variance of the parameters. If this proves true, the more complete jackknife approach should be adopted in operational practice.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and test are measuring. *Applied Measurement in Education, 7,* 255–278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–329.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6), 716–723.*

Allen, N. A., Donoghue, J. R., & Schoeps, T. L. (Eds.). (2001). *The NAEP 1998 technical report* (NCES 2001-452). Washington, DC: National Center for Education Statistics.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279–292.

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

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*(4), 285–291.

Dippo, S., Fay, R. E., & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the American Statistical Association Survey Research Methods Section*. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/1984_023.pdf

Donahue, P. L., Daane, M. C., & Jin, Y. (2005). *The nation's report card: reading 2003* (NCES 2005–453). Washington, DC: U.S. Government Printing Office.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* (pp. 1–19). Philadelphia, PA: Society for Industry and Applied Mathematics.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. In *Proceedings of the Section on Survey Research Methods, American Statistical Association.* Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/papers/1989_033.pdf

Goodman, L. A. *(1974).* Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215–231.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam, the Netherlands: Elsevier.

Haberman, S. J., von Davier, M., & Lee, Y. H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions*. (ETS Research Rep. No. RR-08-45) Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14*(4), 303–334.

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

Lazarsfeld, P. F., and Henry, N. W. (1968), *Latent Structure Analysis*, Boston, MA: Houghton Mifflin.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mazzeo, J., Donoghue, J. R., Li, D., & Johnson, M. (2006). *Marginal estimation in NAEP: Current operational procedures and AM.* Unpublished manuscript.

McCutcheon, A. L. (1987). *Latent class analysis (*Sage University Paper series on quantitative applications in the social sciences, No. 07-064*).* Newbury Park, CA: Sage..

Miller, R. G. (1974). The jackknife: A review. *Biometrika, 61*(1), 1–15.

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

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

Perie, M., Grigg, W., & Donahue, P. (2005). *The nation's report card: Reading 2005* (NCES 2006–451). Washington, DC: U.S. Government Printing Office.

Qian, J., Kaplan, B. A., Johnson, E. G., Krenzke, T., & Rust, K. F. (2001). Weighting procedures and estimation of sampling variance for the national assessment. In N. A. Allen, J. R.

Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report* (NCES 2001-452). Washington, DC: National Center for Education Statistics.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rust, K. F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381–397.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex survey using replication techniques. *Statistical Methods in Medical Research, 5*, 283–310.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

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.

von Davier, M. (2003). *Comparing conditional and marginal direct estimation of subgroup distributions* (ETS Research Rep. No. RR-03-02). Princeton, NJ: ETS.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2007). *Mixture of general diagnostic models* (ETS Research Rep. No. RR-07-32). Princeton, NJ: ETS.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 643–768). Amsterdam: Elsevier.

von Davier, M. & Sinharay, S. (2004). Application of stochastic EM method to latent regression models. (ETS Research Report, RR-04-34). Princeton, NJ: ETS.

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. Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: Elsevier.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389–406.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics, 56*(2), 645–646.

Xu, X. (2007). *Monotone properties of a general diagnostic model* (ETS Research Rep. No. RR-07-25). Princeton, NJ: ETS.

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

Xu, X., & von Davier, M. (2008). *Fitting the structural general diagnostic model to NAEP data* (ETS Research Rep. No. RR-08-27). Princeton, NJ: ETS.

Notes

[1] *CGROUP* uses a Laplace approximation and is designed to be computationally feasible for a test with more than two dimensions (von Davier & Sinharay, 2004).

[2] Namely, reading for literary experience and reading to gain information.

[3] Namely, forming a general understanding, developing interpretation, making reader/text connections, and examining content and structure.

[4] In 2005, the definition of the national sample was changed: it now includes all of the international Department of Defense schools (Perie, Grigg, & Donahue, 2005).

[5] 62 was selected in NAEP operational analysis by design.