Variance estimation for NAEP data using a comprehensive resampling-based approach: An application of cognitive diagnostic models[1]

Chueh-an Hsieh
Michigan State University, East Lansing, Michigan, United States of America

Xueli Xu and Matthias von Davier
Educational Testing Service, Princeton, New Jersey, United States of America

This article presents an application of the jackknifing re-sampling approach (Efron, 1982) to error variance estimation for ability distributions of groups of students, using a multidimensional discrete model for item response data. The data utilized to examine the approach came from the National Assessment of Educational Progress (NAEP). In contrast to the operational approach used in NAEP, where plausible values are generated using the complete sample and are then subjected to a resampling scheme, the proposed approach re-estimated all model parameters for each of the replicate samples during the jackknife. The resampling approach proposed here is therefore a more comprehensive one because of its expected ability to represent the uncertainty due to sampling more appropriately. Results for the comprehensive resampling and re-estimation-based standard errors are presented for estimates of group means, total means, and other statistics used in NAEP for official reporting. Differences in results between the proposed approach and the operational approach are discussed.

1 The authors thank Dan Eignor, Yue Jia, Frank Rijman, and Andreas Oranje for their comments and suggestions, and Kim Fryer for her assistance in copyediting. The authors also thank Steve Isham for providing the data used in this study.

INTRODUCTION

The statistics reported in educational surveys form the basis for secondary analyses and policy research, the outcomes of which guide educational planning. As an ongoing national survey, the National Assessment of Educational Progress (NAEP) is designed to provide national and state information on the academic performance of United States 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 United States Department of Education's National Center for Education Statistics (NCES), and includes a range of surveys and assessments that provide information on students' educational experiences, teachers' characteristics and practices, and school climate.

As is the case with many national surveys, NAEP has adopted a complex sampling design for selecting student participants to 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 non-response and post-stratification). 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 was observed (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001). Another important source of variability of NAEP scores is measurement error. Because items in NAEP assessments are administered according to a partially balanced incomplete block (pBIB) design, each student responds to 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, & Johnson, 2006; Qian et al., 2001).

A major goal of NAEP is to provide estimates of group-level distributions of student proficiencies in the target population as well as in subpopulations of United States 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,[2] 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 in which the parameters of a latent regression model are estimated in the first stage, assuming the item parameters are fixed. This model, with its estimated parameters, is then used to generate a set of plausible values (Mislevy, 1991) that can be considered multiple imputations from the posterior distribution, given students' responses to cognitive

2 CGROUP uses a Laplace approximation and is designed to be computationally feasible for a test with more than two dimensions (Thomas, 1993; von Davier & Sinharay, 2007).

items and background data. These plausible values, in turn, are used to obtain the estimates of interest, for example, group means, standard deviations, percentiles, and other summary statistics. A jackknifing approach based on the single, operational, set of plausible values is adopted in NAEP to obtain estimates of variability for the different statistics of interest.

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 may be 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 one such model that allows different ability variances in different subgroups (Xu & von Davier, 2006, 2008). In addition, the GDM allows a quite parsimonious specification of multiple-group models by utilizing constraints, making it possible to estimate the item parameters and the parameters in the regression models simultaneously. This allows one to utilize the GDM for the required repeated estimation of all model parameters for each of the jackknifing samples in resampling-based variance estimation. In contrast, the current NAEP operation does not allow simultaneous estimation of all model parameters for each jackknifing sample. 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 to compare the variance estimates obtained with the operational results.

GENERAL DIAGNOSTIC MODELS (GDM)

The GDM (von Davier, 2005) contains a large array of psychometric models, such as the latent class analysis (LCA) (Goodman, 1974; Lazarsfeld, & Henry, 1968; McCutcheon, 1987), as well as discrete latent trait models, with pre-specified skill profiles and levels, and multidimensional IRT models (MIRT) (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 latent class analysis, 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 model (2PL-IRT) (Lord & Novick, 1968), 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 applied a compensatory GDM and used the software *mdltm* (von Davier, 2005) to estimate a MIRT model for NAEP data. In addition, we adopted a log-linear smoothing technique to facilitate the estimation of the latent skill space (Xu & von Davier, 2008). Using a log-linear smoothing method allowed us not only to substantially reduce the number of estimated parameters (associated with the latent skill distribution) but also to account for the interrelationship among distinct latent skills. 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).

The Logistic Formulation of a Compensatory GDM

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

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

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 $k = 1,...,K$; $\vec{q}_i = (q_{i1},...,q_{iK})$ 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, some necessary constraints on $\Sigma_k \gamma_{ikc}$ and $\Sigma\beta_{ikc}$ have to be imposed, much like the constraints needed to remove indeterminacy in unidimensional IRT models. Note that a non-zero Q-matrix entry implies that a slope parameter γ_{ikc} is estimated. These slope parameters quantify 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.

Loglinear 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:

$$\log(P_g(a_1, a_2, ...,a_k)) = \mu + \sum_k \beta_{k,g}\,a_k + \sum_k \gamma_k\,a_k^2 + \sum_{i \neq j} \delta_{ij}\,a_i\,a_j \qquad (2)$$

where μ, β_k, γ_k and δ_{ij} are parameters in this log-linear smoothing model, and g is a group index (Haberman, von Davier, & Lee, 2008; Xu & von Davier, 2008). While the GDM and the *mdltm* software allow higher order moments to be estimated, the model in (2) indicates that, in our application to NAEP data, we used only linear and quadratic terms.

SAMPLE AND DATA SOURCES

Data from NAEP 2003 and 2005 fourth-grade reading assessments were used in this study. A representative sample of approximately 191,000 fourth-graders from 7,600 schools was drawn in 2003 by the consortium conducting the NAEP. The operational reporting includes results 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 non-public school student performance. Generally, NAEP reports not only the overall results but also the performance of various subgroups of students, where 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[3] and four reading aspects[4] were specified in the framework of the 2003 reading assessment to evaluate fourth-graders' reading performance, such as population-related means, standard deviations, and 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 2003, the Grade 4 students were given a test booklet consisting of two 25-minute blocks, where reading scales were summarized by three types of questions (i.e., multiple-choice, short constructed-response, and extended constructed-response; see Table 1). 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.

Table 1: The number of items in the NAEP 2003 and 2005 reading assessments

Year	Subscales	Response categories in the item			Total
		Multiple-choice	*Short construct*	*Extended construct*	
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

3 Namely, reading for literary experience and reading to gain information.

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

In similar manner to the 2003 procedure, a nationally representative sample of more than 165,000 fourth-grade students participated in the 2005 assessment. The national results were based on a representative sample of students in both public and non-public schools.[5] The framework used for the NAEP 2005 reading assessment was the same as that used in 2003 (Perie, Grigg, & Donahue, 2005).

VARIANCE ESTIMATION

Estimation of the sampling variability of any statistics should take account of the sample design. In survey practice, the simple random sampling assumption is often violated. Thus, the proper estimation of the sampling variability of a statistic in survey data requires techniques beyond those commonly available in standard packages. There are two commonly used approaches for estimating variances in the analysis of surveys. One is the Taylor series linearization method used to account for complex sample design (Binder, 1983; Li & Oranje, 2007; Williams, 2000); the other is the replication method, which involves recomputing the statistic of interest through use of subsets of data different from but comparable to the original sample and thereby measuring the variance of the parameter estimator (Fay, 1989; Rust, 1985).

In the present study, we applied 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). By permitting fractional weighting of observations, the class of replication methods becomes considerably broader and more flexible (Fay, 1989). By associating replicate weights with the characteristics of the observed sample cases, the replicate weighting approach lends itself particularly well to analysis of data with highly complex design features (Dippo, Fay, & Morganstein, 1984).

NAEP uses a modified BRR, derived from the jackknife procedure (Miller, 1974), to obtain the variance estimate of a statistic. There are 62[6] jackknife samples with different sets of student replicate weights (*SRWTs*). The *SRWTs* are derived from adjustments to the initial base weight. Examples of the adjustments may include non-response, trimming, post-stratification, and the probability of selection for each primary sampling unit (Allen, Donoghue, & Schoeps, 2001). The estimated sampling variance of an estimator, t, is calculated by aggregating these 62 squared differences, $\hat{v}(t) = \sum_{i=1}^{62}(t_i - t)^2$, 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 can refer to the paper by Johnson (1989, p. 315).

5 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).

6 This number is used in NAEP operational analysis.

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; under a multiple-group assumption, different latent skill distributions are allowed for different groups. In this study, the multiple-group variable was defined by race/ethnicity. Four ethnicity groups were distinguished to form the different levels of this variable: White, Black, Hispanic, and Asian/Pacific Islander. As shown in Tables 2 and 5, the results from using the multiple-group assumption, when compared to the results from the single-group assumption, showed better fit in terms of the several fit indices, such as the Bayesian information criterion (BIC; Schwarz, 1978), Akaike's information criterion (AIC; Akaike, 1974), and log-likelihood. Hence, in our study, we compared the results from the multiple-group assumption, such as group means and standard deviations as well as the estimation error of the group mean, with the results from the 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.

NAEP 2003 Reading Assessment for Fourth-grade Students

Table 2 shows the model fit indices under different assumptions. What is evident here is that the multiple-group GDM with race/ethnicity has the better model fit. Thus, this multiple-group GDM will be used and estimates obtained from this model will be compared to the operational results.

Table 2: Model evaluation based on the NAEP 2003 reading assessment

Model	Number 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

Table 3 presents the mean and standard deviation for the race/ethnicity subgroups. As is evident, the results that emerged from using the GDM with the race-group assumption and those that emerged from the NAEP operation have a similar pattern: from high score to low score, the four racial groups have the following order—White, Asian, Hispanic, and African American. We can also observe differences in the means for the subgroups: the differences are relatively large in the subgroups of small sample size (such as in the Asian group). Moreover, the standard deviation estimates obtained from the current approach were smaller for the White and the Asian students and larger for the African American and the Hispanic students.

Table 3: Means and standard deviations for ethnicity subgroups in the 2003 assessment

	GDM				NAEP*			
	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	0.571	0.991	0.495	0.987	0.613	1.030	0.478	1.083

Note: * These results are already on the same scale as those from the GDM runs.

Table 4 shows the comparison between the standard errors associated with the group mean estimates. Here, we can see that the standard errors obtained from using the current approach are slightly larger than those obtained from the operational approach.

Table 4: Standard errors for subgroup mean estimates in the 2003 assessment

White (*N* = 118,061)			
	GDM	*NAEP**	*Ratio of GDM to operation*
Literary	0.007	0.006	1.167
Information	0.008	0.006	1.333
African American (*N* = 35,308)			
	GDM	*NAEP*	*Ratio*
Literary	0.017	0.011	1.545
Information	0.018	0.011	1.636
Hispanic (*N* = 23,839)			
	GDM	*NAEP*	*Ratio*
Literary	0.021	0.016	1.312
Information	0.019	0.017	1.118
Asian (*N* = 8,223)			
	GDM	*NAEP*	*Ratio*
Literary	0.032	0.033	0.970
Information	0.038	0.033	1.151

Note: *These results are not readily available from the NAEP report because the reporting of NAEP is on a scale score metric, not on the θ metric. Instead, these results are derived from the NAEP reporting that involved inversing the transformation.

NAEP 2005 Reading Assessment for Fourth-grade Students

Table 5, which presents the model fit indices, shows that 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. The presentation of the means and the standard deviations for the racial subgroups (Table 6) shows some differences in the ability estimates between the current and the operational approaches with respect to the literary-experience subscale. In addition, the standard deviations for the White and the Asian students tend to be somewhat smaller relative to the current approach than to the operational approach, while the standard deviations for the Hispanic and the African American students are larger for the current than for the operational approach. Note, however, that the patterns of results between the 2003 and 2005 assessments are quite similar.

Table 5: Model evaluation based on the NAEP 2005 reading assessment

Model	Number 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

Table 6: Means and standard deviations for subgroups in the 2005 assessment

	GDM				NAEP			
	Literary subscale		*Information subscale*		*Literary subscale*		*Information subscale*	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
White	0.880	0.872	0.501	1.027	0.910	0.930	0.500	1.028
African American	0.177	1.081	- 0.422	1.067	0.119	0.970	- 0.410	1.049
Hispanic	0.264	1.087	- 0.340	1.112	0.206	1.007	- 0.341	1.102
Asian	0.866	0.921	0.482	1.020	0.917	1.000	0.462	1.114

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

Table 7: Standard errors for subgroup means in the 2005 assessment

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

SUMMARY AND DISCUSSION

The application of the GDM in this article focused not on detecting the skills measured by the NAEP assessment but on improving the estimation of error variances by using a more comprehensive jackknife procedure. The proposed approach accomplishes this by requiring a complete re-estimation of model parameters (item parameters and ability distributions) in jackknife samples, using a multiple-group MIRT model implemented in the GDM framework. Compared to the NAEP operational analysis, where hundreds of background variables are used to extract group ability estimates, the approach chosen for the GDM utilizing only a single grouping variable with four levels is much more parsimonious. In addition, the IRT model used in the GDM analysis does not assume a guessing parameter for multiple-choice items. Given these differences, the results obtained with the two approaches are quite similar. However, the main focus of this paper was to compare the error variance estimation based on one set of (operational) plausible values used in a jackknife scheme, versus a comprehensive re-estimation utilized in the GDM-based jackknife.

Thus, the primary goal of our study was to obtain the estimation error of the subgroup ability means and the standard deviations obtained under the GDM framework. Specifically, we used, in our current procedure, 62 jackknife samples coupled with different sets of weights utilized in the NAEP operational analysis. The results showed that the estimation errors for the ethnicity subgroup means were slightly larger in the proposed approach than were those in the operational. This may be because NAEP operational procedures ignore uncertainty in the item parameters due to sampling.

There are a number of differences between the approach taken using the GDM and that taken using operational procedures. The operational approach assumes normality in the conditional distribution of the latent trait because of the item responses and the large number of background variables (von Davier, 2003). In contrast, the GDM approach does not assume a particular form of the multidimensional ability distributions. Most importantly, the item parameters in the operational analysis are assumed to be fixed and known for the purpose of estimating both the population model and the jackknife replications; our proposed approach re-estimated the item parameters and population distributions for each of the 62 jackknife samples. The capability to re-estimate all parameters used in the GDM enables one to implement a complete jackknife procedure, which results in relatively larger error variance estimates for the group ability means.

Application of the GDM to the NAEP assessment data is not limited to what we have shown in this article.[7] The GDM is able not only to facilitate dimensionality exploration of the NAEP assessment (von Davier, 2005) but also to reduce the number of background covariates when one makes inferences about the group ability estimates. For example, the multiple-group variant of the GDM allows for possible different ability distributions (with potentially different covariance structures) across groups. This heterogeneity of variance structure may reduce the secondary bias in the group mean estimates (Thomas, 2000).

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 not straightforward, a careful judgment process involving content experts and psychometricians is needed during formulation of appropriate models. Moreover, results and inferences must be suitable to be communicated in ways that are useful to stakeholders. For instance, Xu (2007) recently 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. In recent years, there have been calls for more and more subscales and skills to be reported in large-scale surveys. This call leads to models that are parametrically complex, potentially involving multiple, but potentially redundant, subdomains. The principle of parsimony would indicate that the model used for reporting must be complex enough to provide sufficient skill information but still parsimonious enough for the obtained skill information to be non-redundant and reliable (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 (i.e., using recalibrations of item parameters and re-estimated population models) rather than with a jackknife based on imputations from a model using the complete sample needs further investigation. If this was indeed the reason for observing larger variance estimates

7 What is presented here is part of an ongoing program of research geared toward expanding the analysis and reporting alternatives for NAEP.

during use of the GDM rather than of the operation approach, then an inquiry is needed into whether the added portion of variance reflects the true sampling variance of the parameters. If this proves true, then the feasibility of implementing a more complete jackknifing approach into operational analysis procedures should be studied.

References

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. (2001). *The NAEP 1998 technical report* (NCES 2001-452). Washington DC: United States Department of Education, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.

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

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 2*0, 311–329.

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.

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* (pp. 113–121). Alexandria, VA: American Statistical Association.

Donahue, P. L., Daane, M. C., & Jin, Y. (2005). *The Nation's Report Card: Reading 2003* (NCES 2005-453). Washington, DC: United States Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Efron, B. (1982). *The jackknife, the bootstrap and other re-sampling plans*. 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* (pp. 212–217). Alexandria, VA: American Statistical Association.

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: 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*. Princeton, NJ: Educational Testing Service.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14*(4), 303–334.

Lazarsfeld, P. F, & 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: Educational Testing Service.

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

Mazzeo, J., Donoghue, J. R., Li, D., & Johnson, M. (2006). *Marginal estimation in NAEP: Current operational procedures and AM*. Prepared for the National Center for Education Statistics (NCES) under Task 2.2.8 of the NAEP in the New Millennium, Continuity and Innovation contract.

McCutcheon, A. L. (1987). *Latent class analysis: Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage Publications.

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: United States Department of Education, National Center for Education Statistics.

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: United States Department of Education, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*s. Chicago: 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 surveys 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. (1993). *The E-step of the MGROUP EM algorithm* (ETSRR-95-05). Princeton, NJ: Educational Testing Service.

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: Educational Testing Service.

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

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

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

von Davier, M., & Sinharay, S. (2007). An importance sampling EM alogorithm for latent regressional models. *Journal of Educational and Behavioural Statistics, 32*(3), 233–251.

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: 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: Educational Testing Service.

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

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