Chapter 23
Applying the General Diagnostic Model to Proficiency Data from a National Skills Survey

Xueli Xu and Matthias von Davier

Abstract Large-scale educational surveys (including NAEP, TIMSS, PISA) utilize item-response-theory (IRT) calibration together with a latent regression model to make inferences about subgroup ability distributions, including subgroup means, percentiles, as well as standard deviations. It has long been recognized that grouping variables not included in the latent regression model can produce secondary bias in estimates of group differences (Mislevy, RJ, Psychometrika 56:177–196, 1991). To accommodate the ever-increasing number of background variables collected and required for reporting purposes, a principal component analysis based on the background variables (von Davier M, Sinharay S, Oranje A, Beaton AE, The statistical procedures used in national assessment of educational progress: recent developments and future directions. In: Rao CR, Sinharay S (eds) Handbook of statistics: vol. 26. Psychometrics. Elsevier B.V, Amsterdam, pp 1039–1055, 2007; Moran R, Dresher A, Results from NAEP marginal estimation research on multivariate scales. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 2007; Oranje A, Li D, On the role of background variables in large scale survey assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, 2008) is utilized to keep the number of predictors in the latent regression models within a reasonable range. However, even this approach often results in the inclusion of several hundred variables, and it is unknown whether the principal component approach or similar approaches (such as latent-class approaches) are able to generate consistent estimates for individual subgroups (e.g., Wetzel E, Xu X, von Davier M, Educ Psychol Meas 75(5):1–25, 2014). The primary goal of the current study is to provide an exemplary application of diagnostic models for

X. Xu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: xxu@ets.org

M. von Davier
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: mvondavier@nbme.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_23

large-scale-assessment data. Specifically, a latent-class structure is used for covariates while continuing to use IRT models for item responses in the analytic model. Previous applications focused on adult literacy data (von Davier M, Yamamoto K, A class of models for cognitive diagnosis. Paper presented at the 4th Spearman invitational conference, Philadelphia, PA, 2004), as well as large-scale English-language testing programs (von Davier M; A general diagnostic model applied to language testing data (Research report no. RR-05-16). Educational Testing Service, Princeton, 2005, von Davier M, The mixture general diagnostic model. In: Hancock GR, Samuelsen KM (eds) Advances in latent variable mixture models. Information Age Publishing, Charlotte, pp 255–274, 2008), while the current application uses diagnostic modeling approaches on data from NAEP.

23.1 Background and Research Questions

The National Assessment of Educational Progress (NAEP) is often called the "Nation's Report Card" and can be considered the standard of measuring academic progress across the United States for fourth-, eighth-, and 12th- grade students. It covers a wide range of subjects, including reading, mathematics, writing, science, and social science. Every 2 or 4 years, these assessments are administered to nationally representative samples in order to measure trends in academic progress over years. Depending on the assessment subject, the nationally representative samples can have sizes from about 12,000 to about 140,000.

Because NAEP aims to measure the academic progress in policy-relevant subgroups and is mandated not to provide measures at the individual level, a sparse matrix sampling design (Johnson, 1992) is employed to administer NAEP assessments so that individual students take only a portion of entire assessment. For example, for subscales defined in the mathematics framework, students take only about 10–30 items from a pool of 100–150 items, and each student is assigned one of a number of different test forms with a different set of items from the pool. The relatively small number of items within subscales does not provide good ability estimates for individual students, but the aggregation of individual ability distributions is suitable to provide precise estimates of subgroup ability distributions. The methodologies utilized to achieve this were described by von Davier et al. (2007) as well as von Davier and Sinharay (2014). For instance, in order to estimate ability distributions for boys and girls, the item responses as well as the self-reported gender variable needs be included in the model in order to obtain precise estimates for the ability distributions (Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). Modeling approaches that include both grouping variables as well as response variables are commonly referred to as multiple group models. The multiple group model used in NAEP—and other assessments—takes IRT to link the item responses and latent abilities and a latent regression model, in which group differences are regressed on a potentially very large number of background variables.

To facilitate analyses and to enable researchers to answer this demand for inclusion of a large variety of grouping variables, the predictors used in this latent regression model (Mislevy, 1991) are preprocessed by extracting principal components from the bulk of background variables (von Davier et al., 2007; Dresher, 2006; Moran & Dresher, 2007; Oranje & Li, 2008). In the currently operational latent regression procedure used in NAEP, individual subgroup indicators are not used directly. An analysis is conducted to extract principal components (PCs) that explain 90% of the variance of the observed grouping variables. These principal components are then used as the predictors in the latent regression model. This preprocessing raises a question: Is this approach suitable for providing reasonably good estimates of subgroup ability distributions, or does the preprocessing remove some of the between-group differences since the set of predictors used in the model might incompletely reflect differences in grouping variables? It is known that the estimates for subgroups that are not included in the model are usually biased to a certain degree (Mislevy, 1991), but, to our knowledge, the extent to which estimates are biased for subgroups that are only partially represented by means of proxy variables in the form of PCs is unknown. Existing studies were not able to provide definitive answers. For example, Dresher (2006), via a simulation study, found that with a reasonable number of items per students, the latent regression model with PCs (that explains 90% of variance) outperformed the latent regression model with only the subgroup variable of interest in terms of bias and root mean square error. However, Oranje and Li (2008) did not find alarming differences between these two types of models using real data. It is noted that both research studies used NAEP operational software to obtain estimates, assuming that the students share a common covariance structure and their abilities follow a multivariate normal distribution. Their conclusions might change if we use estimation algorithms that allow the assumption of a normal distribution to be relaxed.

The current chapter aims to demonstrate how the estimates of subgroup ability distributions may change when using a different type of conditioning model. For example, we want to obtain estimates of ability distributions of subgroups defined by a background variable using relaxed assumptions. The data analysis uses IRT to calibrate item parameters and takes a conditioning model (either the operational model or some alternative) to estimate the ability distributions for subgroups of interest. To estimate the ability distributions for subgroups, three types of conditioning models were considered and compared (from simplest to most complicated): (a) a model with only the subgroup variable of interest involved, (b) a model with the subgroup variable of interest as well as another important background variable, and (c) a model with preprocessed predictors—in the form of latent class indicator variables. Unlike the approaches used in Wetzel, Xu, and von Davier (2014) where the probabilities of latent classes were used as predictors in the latent regression models, the latent class membership is used in this paper to conduct a multiple-group/multi-dimensional analysis. If the number of latent classes is large enough to sufficiently account for the variation among students, we assume that all three conditioning models will produce approximately the same estimates for

the ability distributions of the subgroup of interest. However, a full model with not-saturated latent classes might incorrectly reflect the group differences. These three conditioning models were analyzed under the general diagnostic model (GDM) framework and compared using NAEP data.

The rest of this chapter is organized as follows: Section 23.2 briefly introduces the GDM, Sect. 23.3 describes the data and procedures used in this study, and Sect. 23.4 shows the results obtained by using the GDM software *mdltm* (von Davier, 2005) for estimation. The last section discusses some of the results and provides further thoughts on the research question.

23.2 The General Diagnostic Model

The GDM (von Davier, 2005) is one of the general frameworks for cognitive diagnostic modeling. As the name suggests, the GDM, as the other cognitive diagnosis models, is mainly developed to diagnose skill levels on finer-grained skills for individual test takers. For example, in the analysis of the well-known fraction subtraction data (Tatsuoka, 1983), the rule space approach, which can be viewed as a deterministic cognitive diagnosis model, was used to make judgments on whether certain skills that are related to fraction calculations are mastered by individual students. Usually, for a test that requires multiple skills, a Q-matrix (Tatsuoka, 1983) is defined based on expert judgments and describes which items require which skills. Quite a few cognitive diagnosis models have been developed in the last two decades, and many of these are described in the first part of this book.

Practically all probabilistic models for cognitive diagnosis can be described as located latent class models (von Davier, 2009). This also applies to the GDM, which expresses the levels for each of the skills as locations on the real line. While this is straightforward for the GDM as it defines a dichotomous or polytomous latent variable for each skill (von Davier, 2005, 2008), even the mastery/non-mastery variable as used in the DINA model (see Chaps. 1 and 7 in this volume) can be defined by two real numbers. This gives a meaning to mastery levels. By using real-valued located latent class, the GDM can easily be extended to more than two ability levels on each of the skills variables. In addition, the GDM bridges the gap between diagnostic models and multidimensional IRT models, and it can be shown that this approach can fit data as well as MIRT models with a multivariate normal ability distribution (Haberman, von Davier, & Lee, 2009). Hence, the GDM can be used to estimate item parameters and latent ability distributions just as commercial IRT estimation programs such as Parscale/or software for MIRT estimation usually do by specifying the skill levels as quadrature points. However, the multivariate ability distribution used in the GDM implementation (Xu & von Davier, 2008a, 2008b) can be estimated freely and, therefore, is more flexible than a (multivariate) normal ability distribution.

23.3 Methodology

23.3.1 Data

Data from a NAEP Grade four reading assessment administered to a national sample was used in the analysis. The data includes 97 items in total and about 140,000 students, each of whom received about 30 items in a balanced incomplete block design (von Davier et al., 2007). The background information we considered for this study includes gender, race/ethnicity, individualized education plan (IEP), limited English proficiency (LEP), free school lunch, location, and computer access at home.

This NAEP instrument measures reading abilities in two content areas: (a) the literary subscale (47 items) and (b) the informational subscale (50 items). By design, the two subscales share no common items.

23.3.2 Analysis Procedure

For each subgroup under consideration (race/ethnicity subgroups, gender groups, as well as school lunch groups), a number of models were estimated: (a) M1: subgroup-based two-dimensional model and (b) M2:latent-class-based two-dimensional model. The models were defined as follows:

M1: Subgroup-based two-dimensional model

In this model, item parameters and subgroup ability distributions are obtained simultaneously by calibrating a two dimensional IRT model (literary and informational subscales as two dimensions) in multiple populations defined by the grouping variables, while restricting item parameter estimates to be the same across subgroups. Under model M1, we estimate a multiple-group model with known assignment of each student to the subgroup of interest. Note that only a single nominal or dichotomous grouping variable is used in these cases, and that ability estimates are based on a Bayesian approach. The *mdltm* software allows expected a-posteriori, or maximum a-posteriori, or imputations based on the posterior distribution. Therefore, for race/ethnicity group comparisons, a model that contains the race/ethnicity grouping variable is appropriate, while for the school-lunch group-based analyses, comparisons of means of estimates between these groups only are appropriate.[1] This implies that any multiple-group analysis is a

[1]This is relevant in cases where Bayesian estimates of ability are used, and the knowledge about grouping, including the differences in ability distributions across groups, is utilized in the estimation. In cases where maximum-likelihood (ML), or bias-corrected ML is used, a multiple group model with item parameter equality will not produce more than trivially different estimates when different grouping variables are used, unless the item parameter estimates are affected by the grouping variables used. Note however, that ML and bias-corrected ML do not reduce measurement

one-time deal: If the grouping variable in the analysis model is not the same as the variable of interest; the estimates obtained from the multiple group model cannot be used for group comparisons. Based on the results reported by Mislevy (1991) and other subsequent publications on the use of this methodology, the gender-based calibration will not be suitable for race/ethnicity group comparisons, as these will likely result in secondary-biased estimates since the analysis model had no information on the race/ethnicity variable. The tables below contain four variants of M1: The first one (M1.1) that includes only the variable of interest, the second (M1.2) includes a grouping variable that is crossed with a second grouping variable (e.g., gender by race/ethnicity with 6 groups fully crossed in the example below), the third (M1.3) includes a grouping variable that is crossed with two other grouping variables (e.g. race/ethnicity by gender and by school lunch eligibility, and a fourth (M1.4) that includes a mismatched grouping variable (e.g., gender variable when deriving race/ethnicity subgroup results).

M2: Latent-class-based two-dimensional model

This model is similar to M1 but differs with respect to how subgroups are defined or, identified. In model M1, the subgroups are defined by observed background variables such as self-reported gender and race/ethnicity. However, in model M2, the subgroups are not assumed to be known, but rather are defined as clusters derived from students' background information. Specifically, the predictors in model M2 are based on the following steps: (a) Fit latent class models to the background variables available in addition to the response data and treat the estimated memberships as if they are observed; (b) then fit a two-dimensional multiple group IRT model to the item responses with groups defined by the estimated class membership.

Under this model, we used 5-latent-classes as groupings for the two-dimensional calibration (M2.1), 10-latent-classes as groupings for a two-dimensional calibration (M2.2) and 50-latent-classes as grouping for a two-dimensional calibration (M2.3). The latent classes based on the background data (e.g., gender, ethnicity, IEP, LEP, etc.) were obtained using the *mdltm* software (von Davier, 2005). The following table lists the levels for each of the background variables included in the latent class approach. The number of potential combinations of the levels of background variables is 1152, which equals the product over the number of levels across these variables given in Table 23.1 below. Each class profile can be represented by 14 parameters, so that there are sufficient degrees of freedom to estimate 10 (140 + 9 parameters) as well as 50 (700 + 49 parameters) latent classes.

This approach is different from assuming a mixture IRT model (e.g., von Davier & Rost, 2007, 2016) which would assume that there are unobserved groups that establish the differences in ability distributions. Instead, the background variables are used in the process of defining populations, but instead of principal components, class membership variables based on a latent class analysis involving all grouping variables of interest are generated.

error due to information about covariates, which is the main reason why background variables are used in latent regression models together with Bayesian ability estimates.

Table 23.1 Background variables used to derive latent classes of the NAEP reading student population

Variable	Levels
Gender	2
Race	6
IEP	2
LEP	2
School lunch	3
Location	4
Computer access	2

23.4 Results

The two models M1 and M2 are compared in terms of subgroup mean and standard deviation (SD) estimates, taking the correct-subgroup two-dimensional calibration (M1.1) as the baseline for comparisons. For example, if the target of inference is for a gender group, we examine the estimates of gender group means and SDs from the multiple group model that contains the gender variable only (correct grouping variable used) to the estimates from the other multiple-group models. The comparisons on estimates for each subgroup have two tables. The first one lists the mean and SD estimates from the different models and the second one shows the difference ratio (other model/M1.1-1).

Tables 23.2, 23.3, 23.4, and 23.5 present the mean and standard deviation estimates and difference ratios for gender, race/ethnicity, and school-lunch-status subgroups. The subgroups comprise the following proportions of the total sample: White students 51%, Black students 14%, Hispanic students 25%, school-lunch-eligible students 52%, and school-lunch not-eligible students 42%.

The following patterns can be discerned:

- As expected, the incorrect-subgroup models produce estimates that are different from the base model M1.1 (e.g., the model M1.4 for race/ethnicity-group inferences and M1.4 for school-lunch-group inferences).
- The model with an interaction including the reporting subgroup of interest produces estimates that are very close to the base model M1.1, such as M1.2 and M1.3 for race/ethnicity-group inference.
- The model with latent classes returns either a reasonably good estimate or inconsistent estimates compared to the correct-subgroup models. For example, the latent class models provide good estimates for White or Hispanic student groups compared with the baseline model, but showed somewhat imprecise estimates for the Black student group. It is unclear why this happens. One potential explanation is that smaller subgroups may not be fully represented in the 10 latent classes that were obtained. The Black student group makes up about 14% of the total sample, which is smaller than the proportions for the other two ethnicity-based subgroups.

Table 23.2 Mean and SD estimates for race/ethnicity subgroups from different conditioning models

	White (51%)				Black (14%)				Hispanic (25%)			
	Literary		Information		Literary		Information		Literary		Information	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
M1.1: Race-group-two-dimensional calibration	0.28	0.90	0.29	0.86	−0.37	0.86	−0.37	0.79	−0.31	0.87	−0.28	0.82
M1.2: Race-gender-crossed-group-two-dimensional calibration	0.28	0.90	0.29	0.85	−0.37	0.86	−0.37	0.79	−0.31	0.87	−0.28	0.82
M1.3: Race-gender-slunch-crossed-group-two-dimensional calibration	0.28	0.89	0.29	0.85	−0.37	0.88	−0.37	0.81	−0.31	0.88	−0.28	0.83
M1.4: Gender-group-two-dimensional calibration	0.22	0.91	0.22	0.87	−0.33	0.90	−0.30	0.85	−0.24	0.91	−0.22	0.86
M2.1: 5-lca-two-dimensional calibration	0.27	0.89	0.28	0.85	−0.35	0.90	−0.33	0.84	−0.29	0.90	−0.27	0.84
M2.2: 10-lca-two-dimensional calibration	0.27	0.89	0.28	0.85	−0.35	0.89	−0.33	0.83	−0.31	0.88	−0.28	0.83
M2.3: 50-lca-two-dimensional calibration	0.28	0.88	0.28	0.84	−0.35	0.90	−0.34	0.84	−0.30	0.91	−0.27	0.86

Table 23.3 The difference ratios for mean and SD estimates when compared to the race/ethnicity-group-two-dimension-model

	White				Black				Hispanic			
	Literary		Information		Literary		Information		Literary		Information	
	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio
M1.1: Race-group-two-dimension calibration	–	–	–	–	–	–	–	–	–	–	–	–
M1.2: Race-gender-crossed-group-two-dimensional calibration	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
M1.3: Race-gender-slunch-crossed-group-two-dimensional calibration	0%	−1%	0%	0%	0%	2%	0%	2%	−1%	2%	0%	1%
M1.4: Gender-group-two-dimensional calibration	−21%	2%	−22%	2%	−11%	4%	−18%	7%	−21%	4%	−23%	4%
M2.1: 5-lca-two-dimensional calibration	−2%	0%	−2%	0%	−8%	5%	−10%	6%	−5%	3%	−5%	2%
M2.2: 10-lca-two-dimensional calibration	−3%	0%	−3%	0%	−7%	4%	−11%	5%	−1%	1%	−2%	1%
M2.3: 50-lca-two-dimensional calibration	0%	−2%	−1%	−1%	−6%	5%	−8%	6%	−2%	5%	−3%	4%

Note: This model is taken as the baseline model

Table 23.4 Comparisons for school lunch subgroup estimates

	School lunch—Yes (52%)				School lunch—No (42%)			
	Literary		Information		Literary		Information	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
M1.1: School lunch-group-two-dimensional calibration	−0.30	0.87	−0.28	0.82	0.40	0.88	0.41	0.83
M1.2: School lunch-LEP-two-dimensional calibration	−0.30	0.88	−0.28	0.82	0.40	0.87	0.41	0.83
M1.3: School lunch-LEP-IEP-two-dimensional calibration	−0.29	0.90	−0.28	0.85	0.40	0.85	0.40	0.81
M1.4: Race-gender-group two-dimensional calibration	−0.26	0.89	−0.24	0.84	0.36	0.89	0.36	0.85
M2.1: 5-lca-two-dimensional calibration								
M2.2: 10-lca-two-dimensional calibration	−0.29	0.88	−0.28	0.82	0.39	0.88	0.40	0.84
M2.3: 50-lca-two-dimensional calibration	−0.29	0.90	−0.27	0.85	0.39	0.85	0.40	0.81

Table 23.5 The differences in mean estimates and the ratios of SD estimates when compared to the school lunch-group-two-dimension-model

	School lunch—Yes				School lunch—No			
	Literary		Information		Literary		Information	
	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio
M1.1: School lunch-group-two-dimension calibration	–	–	–	–	–	–	–	–
M1.2: School lunch-LEP-two-dimensional calibration	0%	1%	−1%	1%	0%	−1%	0%	−1%
M1.3: School lunch-LEP-IEP-two-dimensional calibration	−2%	3%	−2%	4%	−1%	−3%	−2%	−3%
M1.4: Race-gender-group two-dimensional calibration	−13%	2%	−16%	3%	−11%	2%	−12%	2%
M2.1: 5-lca-two-dimensional calibration	−4%	1%	−5%	2%	−3%	0%	−4%	0%
M2.2: 10-lca-two-dimensional calibration	−2%	0%	−3%	0%	−2%	0%	−2%	1%
M2.3: 50-lca-two-dimensional calibration	−3%	3%	−5%	4%	−2%	−3%	−3%	−3%

Note: This model is taken as the baseline model

23.5 Summary

With the increasing scope of policy questions being raised in the context of NAEP, the number of background variables collected to obtain information for reporting purposes increased steadily over past assessment cycles. Educational large-scale survey assessments rely more and more on assumptions made in the latent regression in order to include all available background data. These models may use principal components as done in most operational programs (von Davier & Sinharay, 2014) or latent classes, as proposed by Wetzel et al. (2014) as predictors. Both approaches do not fully reflect the variability in the background data, but rather provide statistical summaries of the associations between the background variables collected in the assessment. The individual subgroup identification is replaced by such data summaries. The study presented in this chapter had the goal to investigate the possible effects of this data reduction. The findings reported above (a) confirm that the estimates for subgroups not included in the analysis models are biased, (b) confirm that the estimates for subgroups that are included in the form of fully crossed interaction models are consistent, and (c) raise concern regarding the use of data summaries (either latent classes or principal components) instead of observed background data. It appears that somewhat inconsistent estimates can result, in particular, if the subgroup information is only incompletely reflected in the statistical summaries that were used as predictors in the latent regression model. This implies that additional research may be needed to straighten this out. For example, the use of latent class analysis for auxiliary background data (Thomas, 2002) such as self-reports on out-of-school activities and educational resources at home together with a direct inclusion of the main reporting variables (gender, ethnicity, LEP, IEP, free school lunch) could be a promising way forward.

Note that the results presented here are limited by the number of background variables used to derive the latent classes. Only seven background variables were used. This approach is not comparable to the number of background variables used in the latent regression models applied in operational practice. A future study might expand to include all available background information to derive the latent classes (e.g., combined with the use of automatic variable selection methods).

References

Dresher, A. (2006, April). *Results from NAEP marginal estimation research*. Presented at the annual meeting of the national council on measurement in education, San Francisco, CA.

Haberman, S., von Davier, M., & Lee, Y.-H. (2009). *Comparison of multidimensional item response models: multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

Johnson, E. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 29*, 95–110.

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristic from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–162.

Moran, R., & Dresher, A. (2007). *Results from NAEP marginal estimation research on multivariate scales.* Paper presented at the annual meeting of the national council on measurement in education, Chicago, IL.

Oranje, A., & Li, D. (2008, April). *On the role of background variables in large scale survey assessments.* Paper presented at the annual meeting of the national council on measurement in education, New York, NY.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*, 33–48. https://doi.org/10.1007/BF02294708

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

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 255–274). Charlotte, NC: Information Age Publishing.

von Davier, M. (2009, March). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement—Interdisciplinary Research and Perspectives, 7*(1), 67–74.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 643–661). Amsterdam, the Netherlands: Elsevier B.V.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Vol. 1: Models* (pp. 393–406). Boca Raton, FL: Chapman and Hall/CRC.

von Davier, M., & Sinharay, S. (2014).. Analytics in international large-scale assessments: Item response theory and population models). In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. E. (2007). The statistical procedures used in national assessment of educational progress: 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 B.V.

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis.* Paper presented at the 4th Spearman invitational conference, Philadelphia, PA.

Wetzel, E., Xu, X., & von Davier, M. (2014). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement, 75*(5), 1–25.

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

Xu, X., & von Davier, M. (2008b). *Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model* (Research Report No. RR-08-35). Princeton, NJ: Educational Testing Service.