An Alternative Way to Model Population Ability Distributions in Large-Scale Educational Surveys

Educational and Psychological
Measurement

2015, Vol. 75(5) 739–763

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414558843

epm.sagepub.com

Eunike Wetzel¹, Xueli Xu², and
Matthias von Davier²

Abstract

In large-scale educational surveys, a latent regression model is used to compensate for the shortage of cognitive information. Conventionally, the covariates in the latent regression model are principal components extracted from background data. This operational method has several important disadvantages, such as the handling of missing data and the high model complexity. The approach introduced here to identify multiple groups that can account for the variation among students is to conduct a latent class analysis (LCA). In the LCA, one or more latent nominal variables are identified that can be used to classify respondents with respect to their background characteristics. These classifications are then introduced as predictors in the latent regression. The primary goal of this study was to explore whether this approach yields similar estimates of group means and standard deviations compared with the operational procedure. The alternative approaches based on LCA differed regarding the number of classes, the items used for the LCA, and whether manifest class membership information or class membership probabilities were used as independent variables in the latent regression. Overall, recovery of the operational approach's group means and standard deviations was very satisfactory for all LCA approaches. Furthermore, the posterior means and standard deviations used to generate plausible values derived from the operational approach and the LCA approaches correlated highly. Thus, incorporating independent variables based on an LCA of background data into the latent regression model appears to be a viable alternative to the operational approach.

¹University of Konstanz, Konstanz, Germany

²Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

Eunike Wetzel, Department of Psychology, University of Konstanz, 78457 Konstanz, Germany.

Email: eunike.wetzel@uni-konstanz.de

Keywords

latent classification, large-scale educational surveys, latent class analysis, latent regression model

Introduction

The use of matrix sampling designs and the focus on group rather than individual scores distinguishes large-scale educational surveys such as the U.S. National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS) from other testing programs. In matrix sampling designs, each student takes only a small subset of all items within one assessment, and different students take different, but overlapping, subsets of items. For example, in the TIMSS 2007 eighth-grade assessment, the total item pool consisted of 429 items (215 mathematics and 214 science items). These items were distributed into 14 mathematics item blocks and 14 science item blocks. The 28 item blocks were combined into 14 booklets which each contained 2 mathematics blocks and 2 science blocks. Each student only took 1 of the 14 booklets to keep the testing burden within reasonable limits (90 minutes in total for the eighth-grade assessment). Consequently, under such a design, the number of items each student responds to is too small to make accurate inferences about individual student ability.

In large-scale educational surveys, a latent regression model (LRM; Mislevy, 1984, 1985) is used to compensate for the shortage of cognitive information and to provide population distributions of latent abilities. Conventionally, the covariates in the LRM are principal components (PCs) extracted from background data. In contrast to the matrix sampling used with items assessing cognitive abilities, each student responds to the same student background questionnaire. For each student, additional background information is collected using responses from the teacher questionnaire and the school questionnaire. The input to the PC analysis is a large number of dummy-coded background data where missing data in background variables are treated as a separate category. A large number of PCs that explain 90% of the variance in the background variables is used as predictors in the LRM. Based on the LRM and item response theory (IRT) models for item responses, plausible values for the latent skill variables, that is, the student proficiencies, are drawn.

These large LRMs used operationally have several disadvantages, for example concerning the handling of missing data. Unlike more recent procedures, methods for estimating regression models often cannot explicitly handle missing data and require each participant to have complete data on all variables of interest. Thus, missing data must be dealt with by excluding cases with missing data in the covariates, imputing missing values, dummy coding all variables, and assigning missing values to additional dummy variables, or using a new set of related variables (i.e., the PCs) that explain the variability in the covariates. Furthermore, the estimation of PCs when there are missing values either requires pairwise deletion of rows or some form of imputation. The operational LRMs also pose a significant computational burden since

there is a very large number of variables (often several thousand dummy variables that are “condensed” into several hundred PCs). For these reasons, an alternative approach that yields similar results as the operational method but that is not subject to the same disadvantages is desirable.

The approach proposed and examined in this article is to identify multiple groups that can account for the variation among students using a latent class analysis (LCA; Lazarsfeld & Henry, 1968) on the background variables. The method of LCA appears ideal for background data, as these are often composed of short scales as well as individual questions that typically yield multidimensional nominal data without the overarching structure of a few dominating latent skill variables. These background data, on the other hand, should be serving as the basis for the definition of groups—or sub-populations—that differ in how they respond to the cognitive test of skills. The primary goal of this article is an exploration of the LCA approach. We examine whether LCA-based latent regressions yield similar estimates of group means and standard deviations for abilities compared with the operational procedure.

The following section describes the operational approach to estimating student proficiencies using a two-stage procedure with a measurement model and an LRM. The subsequent section briefly summarizes research evaluating the operational approach and some alternatives that have been suggested previously and then introduces the proposed alternative approach based on LCA. The penultimate section presents an empirical comparison of the different approaches using eighth-grade TIMSS 2007 mathematics data. The results and their implications are discussed in the final section.

Operational Approach

The estimation process for student proficiencies applied by most large-scale educational surveys involves two stages: (a) a measurement (IRT) model for the observed responses to cognitive items y_i , linking these to a proficiency variable θ_i , and (b) an LRM of the latent proficiency θ_i on background variables x_i . Based on the parameter estimates obtained from these two stages, plausible values for the student group proficiencies can then be generated. In the following, the two stages will be explained in more detail. The notation for the models follows Thomas (2002). For more details on the method, please refer to von Davier, Sinharay, Oranje, and Beaton (2007) or von Davier and Sinharay (2013).

Measurement Model

In the first stage, an item response model is fit to the cognitive response data in order to obtain estimates of item parameters. The cognitive items for the i th student are denoted by $y'_i = (y'_{i1}, \dots, y'_{ip})$ and are partitioned into p latent proficiencies. The binary or ordinal item scores, y_{ijk} , $k = 1, \dots, s_i$, for each y_{ij} are coded as $0, 1, \dots, m_{jk}$ for the k th item measuring the j th proficiency ($j = 1, \dots, p$) where s_j denotes the number of

items on proficiency j . A latent proficiency vector $\theta'_i = (\theta_{i1}, \dots, \theta_{ip})$ is defined for the i th student. Different assessment programs use different IRT models for binary data. The three-parameter logistic (3PL) model (Lord & Novick, 1968) is used by NAEP and TIMSS, while the Program for International Student Assessment (PISA) used the Rasch model (Rasch, 1960) from 2000 to 2012. Other binary models used in international large-scale assessments to model the probability of a correct response are the two-parameter logistic IRT model (Lord & Novick, 1968) and the one-parameter logistic model (Verhelst & Glas, 1995). The 3PL model applied in this study is conditional on x_i, θ_i , and the item parameters $\beta = (\beta_1, \dots, \beta_p), \beta_j = \{\beta_{jk}, k = 1, \dots, s_j\}$ and is defined as

$$\Pr(y_{ijk} = 1 | \theta_i, x_i, \beta) = c_{jk} + (1 - c_{jk}) / [1 + \exp\{-a_{jk}(\theta_{ij} - b_{jk})\}], \quad (1)$$

where $\beta_{jk} = (a_{jk}, b_{jk}, c_{jk})$.

For polytomous items, the generalized partial credit model (Muraki, 1992) can be used,

$$\Pr(y_{ijk} = l | \theta_i, x_i, \beta) = \frac{\exp\left\{\sum_{h=0}^l a_{jk}(\theta_{ij} - b_{jkh})\right\}}{\sum_{q=0}^{m_{jk}} \exp\left\{\sum_{h=0}^q a_{jk}(\theta_{ij} - b_{jkh})\right\}}, \quad l = 0, \dots, m_{jk}, \quad (2)$$

where $\beta_{jk} = (a_{jk}, b_{jkh}, h = 1, \dots, m_{jk})$.

The measurement model invokes the following independence assumptions conditional on the θ_i and β (see Thomas, 2002):

1. The item responses y_i are conditionally independent of the background variables x_i given $\theta'_i = (\theta_{i1}, \dots, \theta_{ip})$
2. The responses of a student to different items are independent conditional on $\theta'_i = (\theta_{i1}, \dots, \theta_{ip})$
3. The responses of different students are independent.
4. The item responses y_{ij} are independent of $\theta_{ij'}$ conditional on $\theta_{ij}, j \neq j'$. That is, item responses depend only on the proficiency to which they are allocated.

With these assumptions, the joint distribution of $y'_i = (y'_{i1}, \dots, y'_{ip})$ conditional on θ_i can be represented as

$$\prod_{j=1}^p f_j(y'_{ij} | \theta_{ij}, \beta_j), \quad (3)$$

where each f_j is the product of the item response probabilities given in Equations (1) and (2). Items that were not administered to a student do not contribute to Equation (3) because they are missing completely at random (Little & Rubin, 1987) by design of the booklet assignment.

Latent Regression Model

In the second stage, an LRM (also called population model or conditioning model), in which item parameters are considered known, is estimated. Depending on how many items are administered in the background questionnaire, the LRM can contain many variables (often more than 200). The inclusion of such a large number of variables is based on the assumption that group estimates for a certain variable will only be unbiased if that variable is included in the population model (Mislevy, Johnson, & Muraki, 1992; Rubin, 1987, 1996). PCs of the background variables $x'_i = (x_{i1}, \dots, x_{iq})$ that account for 90% of the total variance are used in the regression. The LRM is defined as

$$\theta = \Gamma'x + \varepsilon, \quad (4)$$

where the error follows a multivariate normal distribution: $\varepsilon \sim \text{MVN}(0, \Sigma)$. That is, it is assumed that the conditional distribution of the ability variable is $\phi(\theta_i | \Gamma'x_i, \Sigma)$ where ϕ denotes the multivariate normal density. Then we obtain the likelihood of the observed data

$$L(\beta, \Gamma, \Sigma) = \prod_{i=1}^n \int \phi(\theta | \Gamma'x_i, \Sigma) \prod_{j=1}^p f_j(y_{ij} | \theta_{\bullet j}, \beta_j) d\theta, \quad (5)$$

and the posterior distribution of the ability variable can be written as

$$f(\theta | x_i, y_i, \beta, \Gamma, \Sigma) \propto \phi(\theta | \Gamma'x_i, \Sigma) \prod_{j=1}^p f_j(y_{ij} | \theta_{\bullet j}, \beta_j). \quad (6)$$

After estimation of the model parameters in the latent regression (Γ, Σ) a number of multiple imputations (in this context called plausible values) are drawn for each response vector from the posterior distribution given in Equation (6). Details on estimation and the rationale for the use of plausible values can be found in von Davier et al. (2007), von Davier, Gonzalez, and Mislevy (2009), and Mislevy et al. (1992).

Evaluation of the Operational Approach and Alternative Approaches

Several simulation studies have evaluated the operational approach and at the same time compared it with one or more alternative approaches. For instance, Dresher (2006) compared the operational approach with a one-group-at-a-time model (Cohen & Jiang, 1999) in which the analyses are conducted separately for each subgroup of interest and no other background variables are introduced to the model. The operational approach yielded a smaller bias in estimating overall and subgroup mean scores, standard deviations, percentiles, and achievement-level percentages than the one-group-at-a-time model. In absolute terms, the bias was very small (less than or

equal to one scale score point) for the operational approach when the number of items was above 10 and the ratio of sample size to PCs was not too high (optimally around 8 or larger). Furthermore, the operational approach was better able to recover the slightly negatively skewed population distribution than the one-group-at-a-time model. Moran and Dresher (2007) extended this study by applying the comparison to multivariate scales that were both balanced and unbalanced in content (i.e., containing items covering both subscales or not) and found the same general result supporting the operational approach.

Oranje and Ye (2014) also investigated the operational approach and two modifications to it: (a) using the covariance matrix instead of the correlation matrix to extract PCs and (b) using the main reporting variables such as gender, race/ethnicity, and parental education directly, and performing the PC analysis on the residual correlation matrix. For the three model types PCs were extracted that accounted for 90%, 70%, or 50% of the variance. In a simulation study based on NAEP data sets Oranje and Ye found that larger models (90% variance explanation) yielded smaller biases and larger variances than smaller models (50% variance explanation). In total, the third model type, which used the main reporting variables directly, combined with a variance explanation of 90% in the PC analysis yielded the best bias-variance trade-off, though the differences to the operational model were very small.

Another modification to the operational approach that has been proposed involves using an LRM with a reduced number of background variables. Thomas (2002) compared the efficiency of a reduced LRM with the operational (full) model. He found that the efficiency of the estimated means and proportions exceeding cut points were not affected by reducing the number of background variables and suggested that this method could increase testing time for cognitive items and in turn increase the precision of proficiency estimates. However, these results only held for balanced designs and did not deal with secondary biases such as the effect on the variables not included. Furthermore, a reduction of background variables will come at the cost of not being able to report on as many different groups.

In sum, previous research shows that the operational PC approach achieves accurate estimates of student abilities at the group level as long as the scales are not too short or the number of PCs is not too high. Nevertheless, due to the computational complexities of the operational approach described above, alternative approaches should be explored. In light of existing research, the operational approach can be said to yield unbiased estimates, so that alternative approaches can be evaluated with respect to how well they recover results obtained using the operational approach.

Our proposed approach is to use LCA to identify one or more latent nominal variables to classify respondents with respect to statistical dependencies in their background characteristics. These classifications are then introduced as predictors in the latent regression. The approaches based on LCA we examined differed regarding the number of classes, the items used for the LCA, and whether manifest class membership information or class membership probabilities were used as independent variables in the latent regression.

LCA (Lazarsfeld, 1950; Lazarsfeld & Henry, 1968) can be seen as a data reduction method which explains interindividual differences in item response patterns using a reduced number of groups (*latent classes*) represented by the values of a categorical latent variable. In the latent class model for binary variables, the probability of a response pattern r' , following Rost (2004), is defined as

$$p(r') = \sum_{g=1}^G \pi_g \prod_{k=1}^m \pi_{kg}^{r_k} (1 - \pi_{kg})^{1-r_k}, \quad (7)$$

where π_g is a probability parameter for the class size which defines the probability that a randomly selected person will belong to class g . The probability of correctly responding to item k ($k = 1, \dots, m$) in class g is denoted π_{kg} and G is the number of latent classes. The exponent r_k denotes the k th binary component of response pattern vector r' . The model equation for the LCA for polytomous variables is

$$p(r') = \sum_{g=1}^G \pi_g \prod_{k=1}^m \pi_{kr_k g} = \sum_{g=1}^G \pi_g \prod_{k=1}^m P_k(X = r_k | g), \quad (8)$$

where $P_k(X = r_k | g)$ denotes the probability of a response to item k in category r for respondents from class g . Note that $\prod_{k=1}^m P_k(X = r_k | g) = p(r' | g)$ denotes the class-specific probability of a response vector in class g .

Once the class-specific response probabilities and class sizes are estimated, it is possible to compute the class membership probability for each case. The probability that a person with response pattern r' is a member of class g is

$$p(g | r') = \frac{\pi_g p(r' | g)}{\sum_{h=1}^G \pi_h p(r' | h)}. \quad (9)$$

The denominator ensures that class membership probabilities add to unity:

$\sum_{g=1}^G p(g | r') = 1$. For subsequent analyses, persons can manifestly be assigned to the class for which their class membership probability is highest. Respondents with the same observed response vector are assigned to the same latent class. For more detailed information on LCA, the reader is referred to McCutcheon (1987) and Hagenaars and McCutcheon (2002).

In our study, five alternative approaches based on LCA were investigated. These differ with respect to whether manifest class memberships (deterministic LCA) or class membership probabilities (probabilistic LCA) were used and whether background variables were allocated to variable groups based on similar content (variable

Table 1. Summary of Alternative Approaches Based on LCA.

LCA approach	Variables for LCA	Number of latent classes	Type of predictors in LRM
LCA1-d	53 background variables	27	Manifest class membership (coded 0/1)
LCA2-p	53 background variables	27	Class membership probabilities
LCA3-vg	53 background variables distributed into 5 variable groups	5 per variable group (25 in total)	Class membership probabilities
LCA4-vg3a	3 variable groups with between 6 and 17 background variables (high correlations with proficiency)	5 per variable group (15 in total)	Class membership probabilities
LCA5-vg3b	3 variable groups with between 11 and 17 background variables (moderate correlations with proficiency)	5 per variable group (15 in total)	Class membership probabilities

Note. LCA = latent class analysis; LRM = latent regression model.

groups LCA). Table 1 gives a summary of the five approaches. The five approaches are

1. LCA1-d: Variables representing the contrast-coded manifest membership to 1 out of 27 latent classes (deterministic LCA) were used as independent variables in the latent regression.
2. LCA2-p: Class membership probabilities to all 27 latent classes (probabilistic LCA) were used as independent variables in the LRM.
3. LCA3-vg: Variables from the background questionnaire were clustered into five content-related variable groups (vg). Class membership probabilities from separate LCAs for the five variable groups with five latent classes each were used as independent variables in the LRM.
4. LCA4-vg3a: Class membership probabilities from separate LCAs for a set of three out of five of the variable groups in LCA3-vg with five latent classes each were used as independent variables in the LRM. The selection of these three sets was based on a previous analysis of the correlations between background variables and plausible values derived from an LRM estimated according to the operational approach. The three variable sets containing variables that correlated highest with the plausible values were selected for LCA4-vg3a.
5. LCA5-vg3b: Class membership probabilities from separate LCAs for a second set of three out of five of the variable groups in LCA3-vg with five latent classes each were used as independent variables in the LRM. For LCA5-vg3b, the three sets containing variables that showed lower correlations with the plausible values were chosen.

LCA1-d and LCA2-p only differ with respect to whether the independent variables in the LRM were based on manifest class membership information (LCA1-d) or class membership probabilities (LCA2-p). The probabilistic approach contains more information since it takes into account the probability of membership in each of the latent classes whereas the deterministic approach only takes into account the information which of the latent classes the respondent was assigned to. Therefore, for LCA3-vg, LCA4-vg3a, and LCA5-vg3b, only the probabilistic approach was applied.

Note that using a specific number of 27 (or 26, or 28, or 29) latent classes is somewhat irrelevant here. It was important to use a larger number of classes than potentially necessary, even if this is known to involve the threat of local maxima and identification issues. We used 27 classes simply because the set of background variables contains about twice as many observed variables, so we halved the number of observed variables. Vermunt, van Ginkel, van der Ark, and Sijtsma (2008) explain why identification and local maxima are less of a concern if an LCA classification is used as an auxiliary variable to explain dependencies in an imputation model, which is the case here for the latent regression, since it is essentially an imputation model for plausible values.

Comparison of Approaches Using TIMSS 2007 Data

Using eighth-grade TIMSS 2007 mathematics and background data, we investigated whether alternative LRMs based on LCA could recover the results obtained from the operational LRM using PC analysis. Thus, both methods are identical regarding the first stage of estimation (the measurement model) but differ in the way independent variables for the LRMs are derived.

The five LCA approaches will be compared with respect to how well they recover posterior means and standard deviations on the mathematics scales obtained from the operational LRM. Furthermore, correlations between the different LCA approaches and the operational PC approach will be compared with respect to posterior means and plausible values. Last, the size of correlations between posterior means and background variables will be contrasted between the LCA approaches and the operational PC approach.

Method

Samples. We used data from the 57 countries that participated in the TIMSS 2007 eighth-grade assessment. Sample sizes per participating entity ranged from 1,777 (Minnesota, United States; benchmark participant) to 7,377 (United States) with a total N of 246,112.

Instrument. Data on cognitive abilities consisted of the 215 items included in the TIMSS 2007 mathematics content domain. The mathematics domain can be divided into the four subscales *number*, *algebra*, *geometry*, and *data and chance*. For each subscale, posterior means and standard deviations were computed and plausible values were generated.

Fifty-three variables from the TIMSS 2007 eighth-grade student questionnaire were used as background variables. (The 54 variables that correlated highest with the plausible values were selected but one was dropped because of a low response rate.) For three of the LCA approaches (LCA3-vg, LCA4-vg3a, and LCA5-vg3b), the 53 variables were sorted into variable groups based on their content. This resulted in five variable groups, namely demographic information (7 items), home possessions (6), time spent on activities (11), frequency of activities in mathematics (17), and feelings about school and mathematics (12).

Analyses. The analyses were carried out in two steps. First, LCAs were estimated to obtain class membership information. Then, the class membership assignment or the class membership probabilities obtained from the LCAs using Equation (9) were introduced to the LRM. In the following, the analyses are described in more detail.

Latent class analyses. LCAs were estimated using the software for multidimensional discrete latent trait models (MDLTM; von Davier, 2005, 2008). MDLTM applies the expectation–maximization algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977) to obtain marginal maximum likelihood estimates of

parameters. For LCA1-d and LCA2-p, latent classes were formed based on the responses to all 53 background variables. For the LCA approaches based on variable groups (LCA3-vg, LCA4-vg3a, and LCA5-vg3b), separate LCAs were estimated for each variable group. The number of latent classes estimated separately for each country was 27 for LCA1-d and LCA2-p and five latent classes for each variable group for LCA3-vg to LCA5-vg3b. The main reason for choosing these numbers were practical considerations: Instead of iteratively testing different latent class models and choosing the one with the optimal number (as, e.g., in von Davier, Naemi, & Roberts, 2011) we decided to use a single model that contains a comparably large number of classes to ensure that differences between respondents can be captured by the class variable. This is not unlike the operational approach that uses a fixed percentage of explained variance to determine the model instead of comparing the predictive power of different sets of extracted PCs.

Latent regression model. Analyses for the LRM were conducted using DGROUP (Educational Testing Service, 2012). DGROUP has three phases of operation. In the data preparation phase, item parameter data from IRT scaling of the cognitive data in PARSCALE (Muraki & Bock, 2003) or MDLTM (von Davier, 2005, 2008) are used as input to the DGROUP program. Moreover, a data reduction process (SWEEP; Goodnight, 1979) is applied which removes regressors that have little or no variance or show high collinearity with other variables. In the estimation phase, DGROUP provides maximum likelihood estimates of Γ , Σ (see Equation 5) in the LRM. Last, in the imputation phase, DGROUP generates multiple imputations from the latent variable posterior distributions of examinees (plausible values).

The LCA approaches differed regarding the number and type of independent variables that were included in the LRM. The characteristics of the five LCA approaches are summarized in Table 1. For LCA1-d, manifest class membership assignments resulting from the LCAs for each country in MDLTM were contrast-coded into 26 variables which were used as predictors in the LRM. For LCA2-p, 27 variables were used which contained the probability of class membership to each of the 27 latent classes. For LCA3-vg, LCA4-vg3a, and LCA5-vg3b class membership probabilities were used as well. Here, the class membership probabilities obtained in the separate LCAs for each variable group were combined for the LRM. For LCA3-vg, class membership probabilities to all 25 latent classes resulting from the separate LCAs conducted for each of the five variable groups were included. For LCA4-vg3a, only class membership probabilities to the latent classes from the variable groups demographic information, home possessions, and frequency of activities in mathematics were used, and for LCA5-vg3b, only class membership probabilities to the latent classes from the variable groups time spent on activities, frequency of activities in mathematics, and feelings about school and mathematics were used.

In addition to the estimation of the LRM using data from the complete sample of every country, posterior means and standard deviations were also obtained separately for gender subgroups and for subgroups based on the response categories for the item *number of books at home*. These two variables were chosen for the subgroup

analyses because subgroup differences in ability are often found for gender and number of books at home (e.g., for the TIMSS 2007 eighth-grade assessment on average across countries girls had higher mathematics achievement than boys; Mullis et al., 2008). Number of books at home had five response categories: *none or very few*, *one shelf*, *one bookcase*, *two bookcases*, and *three or more bookcases*. The category *none or very few* was used as a reference category for examining subgroup differences.

The five LCA approaches were compared with the operational PC approach regarding their ability to reproduce the posterior means and standard deviations obtained applying the operational approach. To this purpose, differences between the values obtained from the PC approach and the LCA approaches were computed. To analyze whether the PC approach consistently yielded more extreme values than the LCA approaches, the frequency with which five percentiles of the posterior means (5, 25, 50, 75, 95) were larger for the PC approach compared with each LCA approach was computed and tested for significance using a sign test (Dixon & Mood, 1946). The null hypothesis tested by the sign test is that the number of countries where the PC percentile is larger than the LCA percentile should be equal to the number of countries where the PC percentile is smaller than the LCA percentile, or $p(PC > LCA) = p(PC < LCA) = 0.5$. Similarly, the recovery of subgroup results concerning gender and the number of books at home was also investigated by first computing the differences between the posterior means and standard deviations obtained from the PC and LCA approaches and then testing whether subgroup differences were more extreme for the PC or LCA approaches using the sign test.

Furthermore, correlations between the PC approach and the LCA approaches concerning the posterior means and plausible values were investigated. Last, correlations between the posterior means and nine dichotomously scored background variables were computed. Applying the sign test, we analyzed whether the mean of the correlations between posterior means and each of the background variables across the four mathematics subscales was higher for the PC approach compared with the LCA approaches.

Results

Preanalyses. In the data preparation phase of the analyses using variable groups, some independent variables were removed during the SWEEP procedure due to collinearity. Collinearity was determined by the incremental variance explanation by a variable given a set of previously selected variables. If the incremental variance explanation was below .001, two variables were considered collinear (Educational Testing Service, 2012). For LCA3-vg this reduced the number of independent variables based on class membership probabilities from 25 to between 15 and 21 across countries. For LCA4-vg3a, the range was between 9 and 13 independent variables, and for LCA5-vg3b, it was between 10 and 13 independent variables. Between 92 and 135 PCs were used across the 57 countries in the operational approach.

Table 2. Average Difference Between Principal Components and Latent Class Analysis (LCA) Approaches in Posterior Means and Standard Deviations on the Mathematics Subscales Across All Countries.

Method	Scale							
	Number		Geometry		Data and chance		Algebra	
	M	SD	M	SD	M	SD	M	SD
LCA1-d	−0.0010	0.0004	−0.0020	0.0043	−0.0043	−0.0056	0.0003	−0.0005
LCA2-p	−0.0013	0.0011	−0.0022	0.0050	−0.0043	−0.0054	−0.0001	0.0005
LCA3-vg	−0.0009	0.0007	−0.0037	0.0068	−0.0045	−0.0044	−0.0003	0.0009
LCA4-vg3a	−0.0004	−0.0002	−0.0027	0.0053	−0.0047	−0.0045	0.0003	−0.0003
LCA5-vg3b	0.0003	−0.0015	−0.0022	0.0043	−0.0047	−0.0050	0.0016	−0.0023

Posterior Means and Standard Deviations. Table 2 shows the differences between the operational PC approach and each of the five alternative LCA approaches regarding the posterior means and standard deviations on the four TIMSS 2007 mathematics scales averaged across all 57 countries. All differences between the PC approach and the LCA approaches were minuscule, and the LCA approaches did not show noteworthy differences in their ability to recover the posterior means and standard deviations obtained in the PC approach.

For the posterior means on the four mathematics subscales, we furthermore analyzed whether the percentiles for the PC approach were consistently larger than the percentiles for the LCA approaches across the 57 countries using a sign test with $\alpha = .05$. As Table 3 shows, for the fifth percentile, this was the case across all four mathematics subscales. For example, for number, the fifth percentile of posterior means based on PC was larger than the fifth percentile of posterior means based on LCA for between 46 and 52 out of 57 countries. For the 25th and 50th percentiles, the LCA approaches mostly yielded larger values than the PC approach. For the 75th percentile, PC values were larger than LCA values for the subscales geometry and algebra, while for number and data and chance, sign test results were mostly non-significant. For the 95th percentile, PC posterior means were larger than LCA posterior means for geometry, and for two LCA approaches (LCA1-d and LCA5-vg3b) also for number, while data and chance and algebra yielded mainly nonsignificant sign test results.

Aggregating the significant results across subscales for each method, the number of percentiles where PC was larger than the respective LCA method and the reverse was mostly balanced. For instance, for LCA2-p, 13 out of 20 percentiles resulted in significant sign tests, and for 7 of these, the PC approach had larger percentiles for the majority of the countries, while for 6, LCA2-p had larger percentiles for the majority of the countries. Exemplarily, the cumulative density plot in Figure 1 shows the percentiles for posterior means on the mathematics subscale number for Algeria. Overall, the different LCA approaches yielded very similar results.

Table 3. Sign Test Results for Five Percentiles on the Mathematics Subscales.

Scale	Method	Percentile 5			Percentile 25			Percentile 50			Percentile 75			Percentile 95		
		PC > LCA	p sign test	Freq	PC > LCA	p sign test	Freq	PC > LCA	p sign test	Freq	PC > LCA	p sign test	Freq	PC > LCA	p sign test	Freq
Number	LCA1-d	50	<.001	21	.015	.004	19	.004	.004	32	.069	.069	34	.037	.037	34
	LCA2-p	52	<.001	18	.002	.052	24	.052	.104	28	.104	.104	28	.104	.104	28
	LCA3-vg	48	<.001	17	.001	.024	22	.024	.097	27	.097	.097	33	.052	.052	33
	LCA4-vg3a	46	<.001	17	.001	.004	19	.004	.104	28	.104	.104	31	.085	.085	31
	LCA5-vg3b	47	<.001	16	<.001	.069	25	.069	.104	28	.104	.104	37	.008	.008	37
Geometry	LCA1-d	49	<.001	22	.024	.002	18	.002	.002	39	.002	.002	39	.002	.002	39
	LCA2-p	48	<.001	26	.085	.004	19	.004	.015	36	.015	.015	39	.002	.002	39
	LCA3-vg	48	<.001	23	.037	.008	20	.008	.024	35	.024	.024	30	.097	.097	30
	LCA4-vg3a	45	<.001	19	.004	.004	19	.004	.008	37	.008	.008	36	.015	.015	36
	LCA5-vg3b	46	<.001	19	.004	.002	18	.002	.004	38	.004	.004	38	.004	.004	38
Data and chance	LCA1-d	50	<.001	28	.104	<.001	16	<.001	.104	29	.104	.104	25	.069	.069	25
	LCA2-p	52	<.001	31	.085	.004	19	.004	.037	28	.037	.037	23	.037	.037	23
	LCA3-vg	52	<.001	32	.069	.004	19	.004	.015	21	.015	.015	28	.104	.104	28
	LCA4-vg3a	51	<.001	27	.097	.037	23	.037	.085	26	.085	.085	28	.104	.104	28
	LCA5-vg3b	54	<.001	24	.052	.008	20	.008	.024	22	.024	.024	27	.097	.097	27
Algebra	LCA1-d	42	<.001	17	.001	.037	23	.037	.008	37	.008	.008	29	.104	.104	29
	LCA2-p	44	<.001	18	.002	.037	23	.037	.004	38	.004	.004	28	.104	.104	28
	LCA3-vg	48	<.001	20	.008	.069	25	.069	.104	28	.104	.104	28	.104	.104	28
	LCA4-vg3a	45	<.001	18	.002	.037	23	.037	.052	33	.052	.052	29	.104	.104	29
	LCA5-vg3b	49	<.001	17	.001	.104	28	.104	.024	35	.024	.024	28	.104	.104	28

Note. Total number of countries = 57. Freq = frequency; PC = principal component; LCA = latent class analysis.

Figure 1. Cumulative density plot of five percentiles for posterior means on the mathematics subscale number for Algeria.

Note. PC = principal component; LCA = latent class analysis.

Considering the high degree of similarity between the LCA approaches, results concerning the differences between PC and LCA for the LRMs based on subgroups for gender and the number of books at home will only be reported for the LCA approach using five variable groups (LCA3-vg). Average differences in posterior means and standard deviations across all countries between the PC approach and LCA3-vg were larger when the samples were divided into subgroups (see Table 4). However, all average differences were still less than .03.

For gender subgroups and subgroups based on the variable *number of books at home*, we investigated whether the differences between subgroups were larger using the PC approach or the LCA approaches. The results of the sign tests for LCA3-vg, LCA4-vg3a, and LCA5-vg3b are depicted in Table 5. Regarding the girl-boy difference, the PC approach consistently yielded a larger difference for the majority of the 57 countries than any of the three LCA approaches across the four mathematics subscales. For LCA3-vg and LCA4-vg3a the difference between the five categories of the variable number of books at home (using *none or very few* as the reference category) was larger compared with the PC approach for the majority of the countries. For LCA5-vg3b, the reverse is true, which may be attributed to the fact that the variable number of books at home was included in the variable groups used for LCA3-vg and LCA4-vg3a but not LCA5-vg3b. In general, the frequency of a larger LCA difference than PC difference across the 57 countries was highest for the comparison of the lowest two categories (*none or very few, one shelf*), while the difference in frequencies was less pronounced for comparisons of the lowest category *none or very few* with the higher categories *one bookcase, two bookcases, and three or more bookcases*.

Table 4. Average Difference Between PC and LCA3-vg in Posterior Means and Standard Deviations on the Mathematics Subscales Across all Countries for Gender and Number of Books at Home Subgroups.

	Subgroup	Scale							
		Number		Geometry		Data and chance		Algebra	
		M	SD	M	SD	M	SD	M	SD
Gender	Female	-0.0266	0.0021	-0.0111	0.0060	-0.0113	0.0035	0.0090	-0.0009
	Male	0.0252	-0.0027	0.0042	0.0061	0.0025	0.0070	-0.0097	0.0006
Number of books at home	None or very few	0.0166	-0.0135	0.0187	0.0013	0.0191	0.0096	0.0141	-0.0021
	One shelf	-0.0077	0.0015	-0.0152	0.0033	-0.0089	0.0082	-0.0044	-0.0005
	One bookcase	0.0092	-0.0072	0.0062	-0.0051	0.0076	-0.0014	0.0048	-0.0083
	Two bookcases	0.0008	-0.0081	0.0098	-0.0075	0.0103	-0.0035	0.0074	-0.0107
	Three or more bookcases	0.0080	-0.0210	0.0020	-0.0063	0.0136	-0.0035	-0.0069	-0.0052

Table 5. Sign Test Results for Differences Between Gender Subgroups and Number of Books at Home Subgroups.

Scale	LCA3~vg			LCA4~vg3a			LCA5~vg3b		
	Freq	PC > LCA	p sign test	Freq	PC > LCA	p sign test	Freq	PC > LCA	p sign test
Gender subgroups									
Number									
Girl-boy	36		.015	36		.015	39		.002
Geometry	39		.002	37		.008	40		.001
D & C	37		.008	39		.002	40		.001
Algebra	42		<.001	42		<.001	44		<.001
Number of books at home subgroups									
Number									
None-shelf	8		<.001	8		<.001	25		.069
None-case	20		.008	16		<.001	52		<.001
None-2cases	11		<.001	10		<.001	53		<.001
None-3cases	20		.008	20		.008	56		<.001
Geometry									
None-shelf	14		<.001	15		<.001	27		.097
None-case	24		.052	23		.037	50		<.001
None-2cases	26		.085	28		.104	54		<.001
None-3cases	22		.024	22		.024	54		<.001
D & C									
None-shelf	8		<.001	8		<.001	31		.085
None-case	19		.004	20		.008	53		<.001
None-2cases	23		.037	24		.052	54		<.001
None-3cases	29		.104	26		.085	56		<.001
Algebra									
None-shelf	13		<.001	16		<.001	32		.069
None-case	18		.002	20		.008	52		<.001
None-2cases	28		.104	29		.104	55		<.001
None-3cases	19		.004	17		.001	54		<.001

Note. The categories for the variable number of books at home are none or very few, one shelf, one bookcase, two bookcases, and three or more bookcases. The category none or very few was used as the reference category. Total number of countries for the sign test = 57. D & C = data and chance; Freq = frequency; PC = principal component; LCA = latent class analysis.

Table 6. Correlations Between Posterior Means on the TIMSS 2007 Mathematics Subscales From the PC Approach and the LCA Approaches Averaged Across Countries.

		PC			
		Number	Geometry	Data and chance	Algebra
LCA1-d	Number	.978	.937	.896	.907
	Geometry	.945	.964	.909	.927
	Data and chance	.905	.906	.964	.870
	Algebra	.907	.918	.862	.976
LCA2-p	Number	.979	.938	.897	.907
	Geometry	.946	.965	.909	.928
	Data and chance	.906	.907	.965	.870
	Algebra	.908	.919	.863	.977
LCA3-vg	Number	.982	.940	.898	.910
	Geometry	.948	.969	.909	.931
	Data and chance	.908	.909	.970	.873
	Algebra	.910	.921	.867	.980
LCA4-vg3a	Number	.980	.939	.898	.908
	Geometry	.949	.967	.909	.930
	Data and chance	.909	.912	.968	.874
	Algebra	.910	.918	.867	.978
LCA5-vg3b	Number	.978	.937	.894	.907
	Geometry	.948	.965	.907	.930
	Data and chance	.908	.910	.966	.872
	Algebra	.910	.918	.862	.977

Note. All correlations are significant at the .001 level. Correlations between the PC and LCA posterior means within each subscale are in boldface (see diagonal). PC = principal component; LCA = latent class analysis; TIMSS = Trends in International Mathematics and Science Study.

Correlations. Correlations between the posterior means from the operational PC approach and the five LCA approaches on the mathematics subscales averaged across all countries are depicted in Table 6. As can be seen in the main diagonals, correlations between PC and each LCA approach were very high, though not perfect, for all LCA approaches. Across all countries, correlations for posterior means ranged between .823 (between LCA1-d and PC) and .995 (between LCA3-vg, LCA4-vg3a, LCA5-vg3b, and PC). Correlations between the PC approach and the LCA approaches concerning the five plausible values generated in DGROUPE for the first mathematics subscale number are exemplarily shown in Table 7 for Serbia. Correlations did not differ notably between the different LCA-based approaches. Correlations between all plausible values were similarly high. Across all countries, the range of correlations between plausible values for number was .480 to .841.

Last, it was analyzed whether correlations between posterior means and nine background variables were higher for the PC approach compared with LCA3-vg, LCA4-vg3a, and LCA5-vg3b. For each method, correlations were averaged across the four mathematics subscales. For some countries, one or more of the background variables

Table 7. Correlations Between the Five Plausible Values Derived From the PC Approach and the LCA Approaches for the Mathematics Subscale Number for Serbia.

		PC				
		PV1	PV2	PV3	PV4	PV5
LCA1-d	PV1	.773	.779	.765	.784	.773
	PV2	.784	.788	.775	.790	.785
	PV3	.778	.776	.775	.783	.778
	PV4	.778	.784	.768	.796	.777
	PV5	.776	.776	.774	.784	.776
LCA2-p	PV1	.773	.779	.767	.788	.780
	PV2	.784	.788	.771	.791	.785
	PV3	.780	.782	.778	.784	.781
	PV4	.776	.779	.766	.791	.775
	PV5	.777	.783	.773	.783	.770
LCA3-vg	PV1	.786	.785	.776	.791	.781
	PV2	.791	.789	.787	.797	.787
	PV3	.783	.786	.773	.786	.784
	PV4	.777	.779	.769	.782	.772
	PV5	.774	.780	.769	.783	.779
LCA4-vg3a	PV1	.776	.781	.774	.786	.786
	PV2	.787	.783	.772	.793	.785
	PV3	.781	.781	.779	.785	.784
	PV4	.784	.785	.773	.787	.777
	PV5	.775	.787	.781	.792	.782
LCA5-vg3b	PV1	.774	.782	.770	.785	.783
	PV2	.781	.779	.773	.791	.783
	PV3	.782	.781	.777	.789	.785
	PV4	.781	.786	.771	.786	.775
	PV5	.772	.786	.777	.789	.776

Note. All correlations are significant at the .001 level. PV = plausible value; PC = principal component; LCA = latent class analysis.

were constant (i.e., only one response option was endorsed by participants). In those cases, the country was not included in the computation of the sign test.

The results showed that for LCA4-vg3a and LCA5-vg3b, a significant majority of mean correlations across the participating countries were higher for the PC approach for most of the background variables (seven out of nine for both; see Table 8). For LCA3-vg, the number of countries where the PC correlation was higher than the LCA correlation only differed significantly from the number of countries where the LCA correlation was higher than the PC correlation for four background variables. For these four (*born in <country>*, *in school something of mine was stolen*, *in school made to do things by other student*, and *in school was left out of activities*), PC correlations were higher than LCA correlations.

Table 8. Sign Test Results for Mean Correlations Between Posterior Means and Background Variables.

Background variable	LCA3-vg		LCA4-vg3a		LCA5-vg3b		N countries
	Freq LCA	PC > p sign test	Freq LCA	PC > p sign test	Freq > LCA	PC p sign test	
Sex	28	.104	34	.037	42	<.001	57
Born in <country >	42	<.001	44	<.001	55	<.001	56
Home possession computer	27	.102	27	.102	54	<.001	56
Home possession Internet connection	24	.060	17	.001	52	<.001	56
In school something of mine was stolen	38	.001	51	<.001	32	.043	54
In school hit or hurt by other student	25	.093	45	<.001	24	.078	54
In school made to do things by other student	43	<.001	52	<.001	45	<.001	54
In school made fun of or called names	23	.060	47	<.001	24	.078	54
In school was left out of activities	46	<.001	53	<.001	46	<.001	54

Note. Correlations are averaged across the four mathematics subscales. Freq = frequency; PC = principal component; LCA = latent class analysis.

Discussion and Conclusion

The use of PC analysis is the operational method to aggregate data for the LRM in many large-scale educational surveys. This method yields independent predictors that account for a large portion of the variance and achieves accurate estimates of student abilities at the group level (Dresher, 2006; Moran & Dresher, 2007; Oranje & Ye, 2014). However, it is only one of many data reduction methods. Several modifications and alternatives have been suggested previously. For example, Thomas (2002) proposed using a reduced number of background variables, a method that yielded satisfactory results. Another alternative approach is using a one-group-at-a-time model (Cohen & Jiang, 1999) though it could not outperform the operational PC approach (Dresher, 2006; Moran & Dresher, 2007).

In this article, a different alternative approach based on LCA was taken which classified students regarding their background characteristics and used these classifications as predictors in the LRM. This approach is similar to applying LCA directly as an imputation model for categorical data which has been shown to yield good recovery of parameters and standard errors when a sufficient number of latent classes is used (Vermunt et al., 2008). Since the parameters of interest in large-scale educational surveys are student proficiencies, we used LCA as a method of obtaining predictors for the LRM. Minuscule differences and high correlations between the PC approach and the LCA approaches for the posterior means and standard deviations on the mathematics subscales indicate that substituting the operational LRM with a more parsimonious LRM based on either class membership probabilities or manifest membership to latent classes did not lead to a loss of information. Differences between PC and LCA were slightly larger but still negligible when subgroups of participants (gender, number of books at home) were considered.

The fifth percentile of posterior means based on the PC approach was consistently larger than the fifth percentile of posterior means based on the LCA approaches as shown by the sign test. However, for the other four percentiles (25, 50, 75, 95), the frequency with which the PC approach yielded larger values than the LCA approaches was lower, and in many cases, not significantly different from the frequency with which the LCA approaches yielded larger values than the PC approach. The sign test for correlations between posterior means and background variables averaged across the mathematics subscales resulted in nonsignificant differences between the number of countries where the PC correlation was higher and the number of countries where the LCA (based on five variable groups) correlation was higher for most of the nine background variables.

Thus, using a relatively large number of classes in the LCA (15 for LCA4-vg3a and LCA4-vg3b, 25 for LCA3-vg, and 27 for LCA1-d and LCA2-p) yields sufficiently fine-grained classifications to account for the variance in the data. Vermunt et al. (2008) note that in the context of using LCA for the purposes of density estimation, fitting a larger number of classes than necessary (overfitting) is advantageous since sample specific variability can be captured whereas underfitting is problematic because important associations or interactions between variables are ignored. In this

study, the LCA approaches were evaluated with respect to how well they recovered the results from the operational approach since this is the method currently used in most large-scale educational surveys and it has been shown to provide unbiased estimates (e.g., Oranje & Ye, 2014). Future research could apply a simulation study to evaluate both approaches regarding their ability to recover the true population values.

While the operational PC approach can explain slightly more residual variance, the LCA approach is the more efficient data reduction method. For the LCA approaches based on sets of variable groups, some independent variables (consisting of class membership probabilities) were removed in the data preparation phase due to collinearity. These analyses were based on a multiple classifications LCA because the class membership probabilities resulting from separate LCAs for each variable group were combined into one LRM. It is conceivable that some of the classifications were redundant due to overlap in the content of the background variables from different variable groups, for example, demographic information and home possessions. Thus, future applications might attempt to use sets of background variables that differ in content and yet are strongly related to the latent proficiencies.

One disadvantage of the operational approach is that the more PCs are included in the LRM, the more concern is warranted that they lose meaning. The LCA approach proposed in this article classifies respondents into groups that share background characteristics. This idea can also be applied to other contexts such as research in cross-cultural psychology where background characteristics could be used to explain country-level variation (for an approach using cluster analysis, see Georgas & Berry, 1995). Since many background variables assessed in the student questionnaire are known to be related to the proficiencies assessed (e.g., those on home resources such as the number of books at home; Mullis et al., 2008), using classifications of students based on these characteristics is a promising approach. Thus, in contrast to the PCs used in the operational approach, relationships between the background variables are taken into account. The LCA approach is also better able to handle missing data compared with the PC approach because respondents are classified despite having missing data in the background variables. Thus, unlike the PC approach, the LCA approach does not require the exclusion of data or imputation when missing values occur. Moreover, the LCA approach does not require all the background variables to be contrast-coded and, in general, fewer coefficients need to be estimated compared with the PC approach. In consequence, the LCA approach is more parsimonious.

The two-stage procedure applied by both the operational approach and the LCA approach proposed in this study neglect the standard errors of the parameter estimates from the first stage during the second stage. Thus, future research could investigate modifying the LCA approach to a single model in which a regression of the cognitive ability variables on the latent classes is specified. The assumptions embedded in the operational LRM such as (a) common regression coefficients when using PCs to predict the ability across all students and (b) a normally distributed residual with a common covariance matrix across students, impose constraints on student ability distributions. For example, a common covariance matrix might not be

appropriate for students from different (sub)populations and the same linear relationship between PCs and latent variables may not hold across subgroups (Thomas, 2002). Thus, one area of future research could aim at using multiple-group analysis to relax these constraints by allowing free estimation of the ability distribution for multiple groups. The probability-based LCA approach applied here already allows for differing group means, though the residual variance assumption was retained in the DGROUP analyses. Extending this approach to multiple-group analysis, which allows different distributions in different subgroups, would be a promising endeavor.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459. doi:10.1007/BF02293801
- Cohen, J. D., & Jiang, T. (1999). Comparison of partially measured latent traits across nominal subgroups. *Journal of the American Statistical Association*, 94, 1035-1044. doi:10.1080/01621459.1999.10473856
- Dempster, A., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39, 1-38.
- Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41, 557-566. doi:10.1080/01621459.1946.10501898
- Dresher, A. (2006). *Results from NAEP marginal estimation research*. Paper presented at the 2006 annual meeting of the National Council for Measurement in Education (NCME), San Francisco, CA.
- Educational Testing Service. (2012). DGROUP [computer software]. Princeton, NJ: Author.
- Georgas, J., & Berry, J. W. (1995). An ecocultural taxonomy for cross-cultural psychology. *Cross-Cultural Research*, 29, 121-157. doi:10.1177/106939719502900202
- Goodnight, J. H. (1979). A tutorial on the SWEEP operator. *The American Statistician*, 33, 149-158. doi:10.2307/2683825
- Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, England: Cambridge University Press.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.

- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381. doi:10.1007/BF02306026
- Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993-997. doi:10.1080/01621459.1985.10478215
- Mislevy, R. J., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, 17, 131-154. doi:10.3102/10769986017002131
- Moran, R., & Dresher, A. R. (2007). *Results from NAEP marginal estimation research on multivariate scales*. Paper presented at the annual meeting of the National Council on Measurement in Education in Chicago, IL.
- Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., & . . . Galia, J. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. doi:10.1177/014662169201600206
- Muraki, E., & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Skokie, IL: Scientific Software International.
- Oranje, A., & Ye, L. (2014). Population model size, bias and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 203-228). Boca Raton, FL: CRC Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Rost, J. (2004). *Lehrbuch Testtheorie—Testkonstruktion* [Textbook test theory—test construction] (2nd ed.). Bern, Switzerland: Verlag Hans Huber.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, 67, 33-48. doi:10.1007/BF02294708
- Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 215-238). New York, NY: Springer.
- Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Social Methodology*, 38, 369-397.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS 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. Samuelson (Eds.), *Latent variable mixture models* (pp. 255-274). Charlotte, NC: Information Age.
- von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and methodologies in large scale assessments (Vol. 2, pp. 9-36)*. Hamburg, Germany: IERI Institute.
- von Davier, M., Naemi, B., & Roberts, R. (2011, April). *Comparing typological and factorial models for personality scales and survey data*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- von Davier, M., & Sinharay, S. (2013). 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* (pp. 1039-1055). Amsterdam, Netherlands: North-Holland.