Article

Journal of Educational and Behavioral Statistics
2017, Vol. 42, No. 5, pp. 591–616
DOI: 10.3102/1076998617700598

Bayesian Estimation of Multivariate Latent Regression Models: Gauss Versus Laplace

Steven Andrew Culpepper
Trevor Park
University of Illinois at Urbana-Champaign

A latent multivariate regression model is developed that employs a generalized asymmetric Laplace (GAL) prior distribution for regression coefficients. The model is designed for high-dimensional applications where an approximate sparsity condition is satisfied, such that many regression coefficients are near zero after accounting for all the model predictors. The model is applicable to large-scale assessments such as the National Assessment of Educational Progress (NAEP), which includes hundreds of student, teacher, and school predictors of latent achievement. Monte Carlo evidence suggests that employing the GAL prior provides more precise estimation of coefficients that equal zero in comparison to a multivariate normal (MVN) prior, which translates to more accurate model selection. Furthermore, the GAL yielded less biased estimates of regression coefficients in smaller samples. The developed model is applied to mathematics achievement data from the 2011 NAEP for 175,200 eighth graders. The GAL and MVN NAEP estimates were similar, but the GAL was more parsimonious by selecting 12 fewer (i.e., 83 of the 148) variable groups. There were noticeable differences between estimates computed with a GAL prior and plausible value regressions with the AM software (beta version 0.06.00). Implications of the results are discussed for test developers and applied researchers.

Keywords: *multivariate regression; Bayesian Lasso; National Assessment of Educational Progress; multivariate generalized asymmetric Laplace distribution; probit model*

Policymakers, applied education researchers, and practitioners rely upon large-scale testing programs, such as the National Assessment of Educational Progress (NAEP), to provide timely data on the status of what America's students know and can do in various subject areas. Large-scale testing programs produce large, nationally representative data sets of student achievement with hundreds of student, teacher, and school administrator survey responses as predictor variables. Accordingly, the analysis of NAEP data provides applied researchers the opportunity to test hypotheses regarding the relationship between student achievement and hundreds of background variables. The fundamental goal of such investigations is to distinguish between those variables that predict achievement and those that do not,

which is referred to as model selection in the statistics literature. Applied researchers typically use readily available software (e.g., the AM software; American Institutes for Research, 2002–2012) to conduct inferences. However, current methods and software are not designed to conduct model selection (i.e., to separate significant and insignificant variables). Instead, the application of current regression procedures may distort researchers' inferences about model selection and thereby impact the process of theory development and testing.

The methodology for scaling and estimating subgroup performance in large-scale testing is well established (see, e.g., Mislevy, Johnson, & Muraki, 1992; Rubin, 1987; Thomas, 2000; Thomas & Gan, 1997). This study focuses on improving statistical inference for model selection in large-scale testing. This article offers at least three contributions to the literature. First, a high-dimensional Bayesian regression model is developed for latent, multivariate outcomes. This model aids in selecting a smaller set of predictors by incorporating a prior distribution for coefficients that is peaked near zero. The developed model has the potential to achieve greater parsimony and approximate model selection when estimating the relationship between hundreds of possibly highly correlated variables and latent achievement. The methodology developed in this study may offer educators, test developers, and policymakers more accurate information regarding which student, teacher, and school variables relate to achievement.

Second, this study extends Bayesian Lasso regression methodology (e.g., Kyung, Gill, Ghosh, & Casella, 2010; Park & Casella, 2008) to multivariate contexts with latent dependent variables (e.g., NAEP student achievement). Prior research developed Bayesian (M. S. Johnson, 2002; M. S. Johnson & Jenkins, 2004; M. S. Johnson & Sinharay, 2015) and non-Bayesian (e.g., Sinharay & von Davier, 2005; von Davier & Sinharay, 2004; von Davier & Sinharay, 2007; von Davier & Sinharay, 2010) methods for estimating the association between background variables and latent achievement. Existing Bayesian methods use a multivariate normal (MVN), or Gaussian, prior for regression coefficients. An MVN prior poorly matches a situation of sparsity, in which there are a few high-magnitude regression coefficients and many others that are near zero after accounting for all the model predictors. One consequence is that employing an MVN prior may not accurately recover the sparse structure (i.e., the insignificant variables) in the multivariate regression coefficients. The developed model employs a multivariate generalized asymmetric Laplace (GAL) prior (Kozubowski, Podgórski, & Rychlik, 2013) for regression coefficients. The GAL distribution is a multivariate generalization of the Laplace distribution, which is peaked near zero and has heavier tails than the MVN prior. The GAL and MVN priors are both expected to provide similar estimates as the sample size increases. The GAL prior is more peaked near the origin and has heavier tails than the MVN, which reduces shrinkage of larger coefficients. Also, results from a Monte Carlo study provide evidence the GAL prior yields more accurate model selection in larger samples.

Third, this article develops a Gibbs sampler for estimating model parameters. The developed algorithm is disseminated as an R package to provide researchers and practitioners with access to the high-dimensional multivariate regression methodology. The Monte Carlo Markov chain (MCMC) routine was written with C++ and Rcpp Armadillo (Eddelbuettel, 2013), and an R package entitled "Latent Multivariate Bayesian Lasso Regression" (lmblr) will be made available on the Comprehensive R Archive Network (CRAN) to assist future research efforts.

The remainder of the article includes three sections. The first reviews frequentist and Bayesian Lasso regression and summarizes large-scale testing methodology. The second section presents a Bayesian formulation for the multivariate latent regression model. In particular, it describes a formulation for jointly estimating item parameters and multivariate regression coefficients with a GAL prior. The third section discusses Monte Carlo simulation results comparing the performance of the GAL and MVN prior for estimating multivariate regression coefficients and performing model selection. The fourth section presents an application of the developed model to the 2011 NAEP mathematics achievement data. Note the GAL results are compared both to the MVN and to estimates using plausible values (PVs) based upon the AM software. The last section provides discussion and concluding remarks.

Literature Review

This section summarizes prior research. The first subsection provides an overview of frequentist and Bayesian Lasso regression and the second subsection reviews existing large-scale testing methodology.

Overview of Frequentist and Bayesian Lasso Regression

Researchers are increasingly interested in testing hypotheses and making predictions in "big-data" contexts with hundreds of predictor variables. In the statistics literature, the least absolute shrinkage and selection operator (Lasso) is a common approach for estimating parameters of high-dimensional regression models and performing model selection (Tibshirani, 1996). For instance, in the context of large-scale educational testing, let $\mathbf{x}_i$ be a vector of V predictors for student i ($i = 1, \ldots, N$) that includes demographic variables and questionnaire responses by students, teachers, and school administrators. Also, define $\mathbf{X}' = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ as a $V \times N$ matrix of predictors for the entire sample of students, $\mathbf{y}$ as an N dimensional vector of dependent variables, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_V)'$ as a vector of regression coefficients. The Lasso was developed to identify relevant predictors, that is, to perform variable selection along with parameter estimation. The Lasso estimate can be described as the minimizer of an L_1-penalized residual sum of squares:

$$\widehat{\boldsymbol{\gamma}} = \underset{\gamma}{\text{argmin}}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\gamma})'(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\gamma}) + \lambda \sum_{v=1}^{V} |\boldsymbol{\gamma}_v|, \tag{1}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_N$, $\bar{y}$ is the sample mean, $\mathbf{1}_N$ is an N dimensional vector of ones, and λ is the Lasso parameter. Choosing larger values of λ gives greater weight to the L_1 penalty function, and elements of $\boldsymbol{\gamma}$ are accordingly shrunken toward zero.

One potential criticism of the Lasso is that its justification seems to assume sparsity or that sparsity must present a problem for the Lasso to be useful. However, as Hastie, Tibshirani, and Wainwright (2015) observe in the context of large-scale regressions, one should "[u]se a procedure that does well in sparse problems, since no procedure does well in dense problems" (p. 2). Sparsity represents an opportunity to be exploited when present.

One limitation of the Lasso as identified by Yuan and Lin (2006) is the treatment of dummy variables for grouping factors. For instance, it is commonplace to include demographic grouping factors such as race/ethnicity using dummy variables. Examples of other "groups" of variables in large-scale assessments include dummy variables used to code school location, teacher certification type, and English-language proficiency status. The traditional Lasso penalizes each variable within a grouping factor rather than penalizing the grouping factor as a whole. One consequence is that the traditional Lasso could select a subset of dummy variables within a grouping factor. The problem with the way the Lasso treats grouping variables is that the interpretation of dummy variable coefficients changes if any one dummy variable is omitted for a group. In contrast, the group Lasso considers all dummy variables corresponding to a factor as a single group to ensure that decisions about statistical significance and variable selection are made for the entire group of dummy variables.

Accordingly, partition $\mathbf{X}$ into $g = 1, \ldots, G$ groups of variables such that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)' = (\mathbf{X}_1, \ldots, \mathbf{X}_G)$ where $\mathbf{X}_g$ is a $N \times V_g$ matrix and $V = \sum_{g=1}^{G} V_g$. Similarly, partition the regression coefficients as $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1{}', \ldots, \boldsymbol{\gamma}_G{}')$ where $\boldsymbol{\gamma}_g$ is a V_g dimensional vector. Yuan and Lin (2006) define the group Lasso estimate as

$$\widehat{\boldsymbol{\gamma}} = \underset{\gamma}{\text{argmin}} \left(\tilde{\mathbf{y}} - \sum_{g=1}^{G} \mathbf{X}_g\boldsymbol{\gamma}_g\right)' \left(\tilde{\mathbf{y}} - \sum_{g=1}^{G} \mathbf{X}_g\boldsymbol{\gamma}_g\right) + \lambda \sum_{g=1}^{G} \sqrt{V_g\boldsymbol{\gamma}'_g\boldsymbol{\gamma}_g}. \tag{2}$$

Equation 2 shows that the Lasso penalty is applied for each group, so that shrinkage of regression coefficients occurs at the group level.

Prior research also considered Bayesian versions of Lasso regression (e.g., Hans, 2009; Kyung et al., 2010; Park & Casella, 2008). Park and Casella (2008) discussed a univariate Bayesian Lasso and Kyung et al. (2010) created the Bayesian group Lasso to handle groups of variables similar to Yuan and Lin (2006). For example, Park and Casella (2008) proposed the following model:

$$\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_N(\mu \mathbf{1}_N + \mathbf{X}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_N), \tag{3}$$

$$\boldsymbol{\gamma}|\sigma^2, \tau_1^2, \ldots, \tau_V^2 \sim \mathcal{N}_V\left[\mathbf{0}_V, \sigma^2 \text{diag}\left(\tau_1^2, \ldots, \tau_V^2\right)\right], \tag{4}$$

$$\tau_v^2|\lambda^2 \sim \text{Exponential}\left(\frac{\lambda^2}{2}\right), \tag{5}$$

$$\sigma^2 \sim \text{inverse gamma}(a, b), \tag{6}$$

$$\lambda^2 \sim \text{gamma}(r, \delta). \tag{7}$$

Equations 3 and 6 include the traditional normal theory Bayesian regression model for $\mathbf{y}$ and a conjugate inverse gamma prior for σ^2. The novelty of the formulation resides with Equations 4 and 5. In particular, integrating out $\tau_1^2, \ldots, \tau_V^2$ from the prior distribution $p(\boldsymbol{\gamma}, \tau_1^2, \ldots, \tau_V^2|\sigma^2)$ implies that the prior distribution of $\gamma_1, \ldots, \gamma_V|\sigma^2, \lambda^2$ is independent and identically distributed (i.i.d.) Laplace,

$$p(\boldsymbol{\gamma}|\sigma^2, \lambda^2) = \frac{1}{2^V}\left(\frac{\lambda^2}{\sigma^2}\right)^{V/2} \prod_{v=1}^{V} e^{-|\gamma_v|\sqrt{\frac{\lambda^2}{\sigma^2}}}. \tag{8}$$

It is well known that the prior for γ in Equation 8 is more peaked at zero, and Park and Casella (2008) provided an example showing that the Bayesian Lasso shrinks coefficients more than ridge regression, but less than the ordinary Lasso. Furthermore, a benefit of employing a fully Bayesian formulation of the Lasso is that the posterior distribution of λ^2 is partly determined based upon the data, so the algorithm shrinks regression parameters in a self-tuning manner. In fact, Park and Casella provided an example where using a diffuse prior for λ^2 yielded comparable results with marginal maximum likelihood estimation.

Overview of Large-Scale Testing Methodology

Large-scale testing programs such as NAEP estimate latent achievement with three essential steps. First, students are administered a random sample of items in a given content area to reduce burden and limit total test time (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; E. G. Johnson, 1992; Mislevy et al., 1992; Zwick, 1987). Let $k = 1, \ldots, K$ indicate dimensions or proficiency content areas (i.e., the number of latent dependent variables); let $j = 1, \ldots, J_k$ index the number of items in content area k and denote the total number of items by $J = \sum_k J_k$. Let the observed item responses for individual i in content area k be denoted by $\mathbf{y}_{ik}$ and let $\mathbf{y}_i' = (\mathbf{y}_{i1}', \ldots, \mathbf{y}_{iK}')$ be an observed vector of item responses across K content areas. Random assignment of items implies a subset of all *ijk* combinations appear in the data. For example, $\mathbf{y}_{ik}$ includes values of $(y_{i1k}, \ldots, y_{iJ_k k})$ for which $j \in \mathcal{T}_{ik}$ where $\mathcal{T}_{ik}$ denotes the set of tasks/items given to student i from content area k.

Similarly, let $\mathbf{y}_k^j$ include values of $(y_{1jk}, \ldots, y_{Njk})$ for which $i \in \mathcal{S}_{jk}$ where $\mathcal{S}_{jk}$ is the set of students who were assigned item j from content area k.

Second, large-scale testing programs estimate item parameters, $\boldsymbol{\Omega}$, by marginalizing over the latent variables. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})'$ be a vector of latent variables underlying the observed $\mathbf{y}_i$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)'$ is an $N \times K$ matrix of latent achievement scores. Programs such as NAEP assume $\boldsymbol{\theta}_i \sim \mathcal{N}_K(\mathbf{0}_K, \mathbf{I}_K)$ (where $\mathbf{0}_K$ is a vector of zeros and $\mathbf{I}_K$ is an identity matrix) and estimate item parameters, $\widehat{\boldsymbol{\Omega}}$, using marginal maximum likelihood.

Third, NAEP uses results from a multivariate regression and a multiple imputation procedure (Rubin, 1987; Thomas & Gan, 1997) to generate PVs with a predictor matrix $\mathbf{X}$, which includes hundreds of columns consisting of student, teacher, and school administrator survey responses. In order to avoid numerical instability due to multicollinearity when generating PVs, the dimensionality of $\mathbf{X}$ is first reduced by retaining the number of principal components needed to account for at least 90% of the total observed variance in $\mathbf{X}$. The subsequent PVs are then summarized using existing software (e.g., the AM software; American Institutes for Research, 2002–2012).

Bayesian Model Formulation

This section discusses the Bayesian formulation of the IRT measurement model and the developed structural model.

Bayesian Formulation of IRT Measurement Model

NAEP uses a collection of IRT models for multiple choice items, constructed responses, and essays. Observed item response y_{ijk} has $M_{jk} \geq 2$ values scored from $m_{jk} = 0, 1, \ldots, M_{jk} - 1$ depending upon whether the item is scored as correct/incorrect, given partial credit, or a rating. The probability of observing responses by student i for items in content area k given $\boldsymbol{\theta}_i$ and $\boldsymbol{\Omega}$ is,

$$p(\mathbf{y}_{ik}|\theta_{ik}, \boldsymbol{\Omega}_k) = \prod_{j \in \mathcal{T}_{ik}} p(y_{ijk}|\theta_{ik}, \boldsymbol{\Omega}_{jk}), \tag{9}$$

where $\boldsymbol{\Omega}_{jk}$ is a row vector of item parameters for item j in content area k and $\boldsymbol{\Omega}_k = (\boldsymbol{\Omega}_{1k}, \ldots, \boldsymbol{\Omega}_{J_k k})$ denotes a row vector of item parameters from content area k. The probability of student i's responses to items across K content areas is,

$$p(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\Omega}) = \prod_{k=1}^{K} p(\mathbf{y}_{ik}|\theta_{ik}, \boldsymbol{\Omega}_k) = \prod_{k=1}^{K} \prod_{j \in \mathcal{T}_{ik}} p(y_{ijk}|\theta_{ik}, \boldsymbol{\Omega}_{jk}), \tag{10}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ is a full row vector of item parameters.

Prior research explored Bayesian IRT models (e.g., Albert, 1992; Albert & Chib, 1993; Fox, 2010; Fox & Glas, 2001; Patz & Junker, 1999). Figure 1 includes a directed acyclic graph of the model developed in this article. Figure

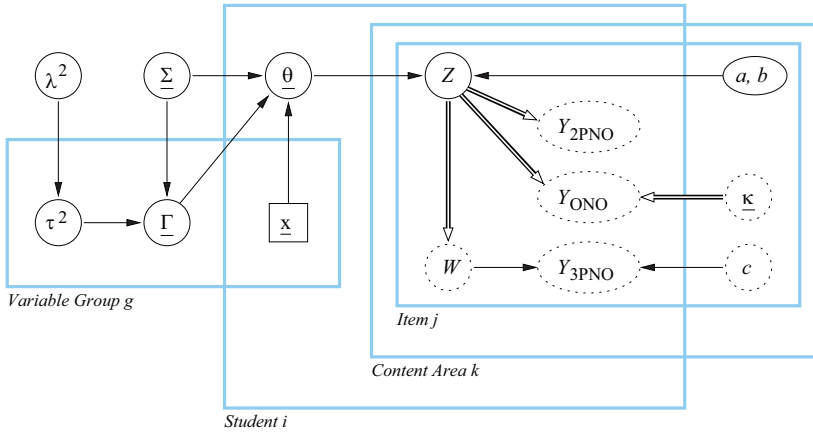

FIGURE 1. *Directed acyclic graph of the measurement and structural models for multivariate continuous latent variables. Circles and ellipses correspond to random variables and squares indicate fixed variables. Dotted circles and ellipses denote variables whose presence depends upon the item type. Solid arrows indicate stochastic dependence whereas hollow lines/arrows denote deterministic dependence. The rectangular plates enclose variables with a common subscript. Underlined quantities are vectors or matrices.*

1 demonstrates that three normal-ogive IRT models are employed for different response types (e.g., multiple choice or constructed response) and scoring procedures (i.e., correct/incorrect or partial credit). It is important to note that normal-ogive measurement models are employed rather than the operational models used in practice, which include the two-parameter logistic (2PL) and three-parameter logistic (3PL) models and the generalized partial credit model (Muraki, 1992). The normal-ogive models are employed for several reasons: (1) availability of Gibbs samplers for individual and item parameters, (2) the choice of "$D = 1.7$" rescales the operational 2PL and 3PL models to resemble normal-ogive item response functions (IRF), and (3) Maydeu-Olivares, Drasgow, and Mead (1994) provide evidence of comparable fit for ordinal models with the same number of parameters. The following subsections discuss the employed IRT measurement models and prior distributions for the item parameters. The full conditional distributions for the IRT model parameters are included in Appendix A, available in the online version of the article.

Two-parameter normal-ogive (2PNO) model. The 2PNO model is used to model dichotomous responses (e.g., variables scored as 0 if incorrect and 1 if correct). The IRF for the 2PNO model is $P(y_{ijk} = 1|\theta_{ik}, \mathbf{\Omega}_{jk}) = \Phi(\eta_{ijk})$ where $\Phi(\cdot)$ is the standard normal distribution function, $\eta_{ijk} = a_{jk}\theta_{ik} - b_{jk}$, and $\mathbf{\Omega}_{jk} = (a_{jk}, b_{jk})$; a_{jk} is an item discrimination parameter that captures the relationship between the

latent variable and item responses; and b_{jk} is an intercept or item threshold parameter. A latent-variable formulation for the 2PNO model is

$$y_{ijk} = \mathcal{I}(Z_{ijk} > 0),$$

$$Z_{ijk}|\theta_{ik}, a_{jk}, b_{jk} \sim \text{independent } N(\eta_{ijk}, 1), \tag{11}$$

where Z_{ijk} is a normally distributed auxiliary variable (Albert & Chib, 1993).

Three-parameter normal-ogive (3PNO) model. The 3PNO model is an extension of the 2PNO that allows for a nonzero probability of guessing. The 3PNO IRF is $P(y_{ijk} = 1|\theta_{ik}, \boldsymbol{\Omega}_{jk}) = c_{jk} + (1 - c_{jk})\Phi(\eta_{ijk})$ where $\boldsymbol{\Omega}_{jk} = (a_{jk}, b_{jk}, c_{jk})$ and c_{jk} is a guessing parameter. Béguin and Glas (2001) extended the 2PNO formulation by introducing a dichotomous augmented variable W_{ijk} for the 3PNO,

$$y_{ijk}|W_{ijk}, c_{jk} \sim \text{independent Bernoulli}\,(c_{jk} + W_{ijk}(1 - c_{jk})), \tag{12}$$

$$W_{ijk} = \mathcal{I}(Z_{ijk} > 0). \tag{13}$$

Ordinal normal-ogive (ONO) model. The ONO is implemented for items that are scored as partial credit or on a rating scale. The IRF for the ONO is $P(y_{ijk} \leq m|\theta_{ik}, \boldsymbol{\Omega}_{jk}) = \Phi(\kappa_{jk,m+1} - \eta_{ijk})$ where $\boldsymbol{\Omega}_{jk} = (a_{jk}, b_{jk}, \boldsymbol{\kappa}_{jk})$ and the row vector of category thresholds is defined as $\boldsymbol{\kappa}_{jk} = (\kappa_{jk0}, \kappa_{jk1}, \ldots, \kappa_{jkM_{jk}})$. The thresholds satisfy the following identifiability constraints: $\kappa_{jk0} = -\infty, \kappa_{jk1} = 0 < \kappa_{jk2} < \cdots < \kappa_{jkM_{jk}} = \infty$. A Bayesian formulation for the ONO model (Albert & Chib, 1993; Cowles, 1996) is

$$y_{ijk} = m \quad \text{if} \quad \kappa_{jkm} < Z_{ijk} \leq \kappa_{jk,m+1}, \tag{14}$$

where κ_{jkm} denotes the $(m + 1)$th threshold for item j in content area k that divides the latent distribution into M_{jk} observed categories.

Prior distributions for item parameters. The following independent priors are employed for item parameters of the 2PNO, 3PNO, and ONO models:

$$\xi_{jk} = (a_{jk}, b_{jk})' \sim \text{iid}\,\mathcal{N}_2(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)\mathcal{I}(a_{jk} > 0), \tag{15}$$

$$c_{jk} \sim \text{iid Beta}\,(\alpha_c, \beta_c), \tag{16}$$

$$p(\boldsymbol{\kappa}) = \prod_{jk} p(\boldsymbol{\kappa}_{jk}) = \prod_{jk} p(\kappa_{jk0}, \kappa_{jk1}, \ldots, \kappa_{jkM_{jk}}), \tag{17}$$

$$p(\kappa_{jk0}, \kappa_{jk1}, \ldots, \kappa_{jkM_{jk}}) \propto \mathcal{I}(\kappa_{jk0} = -\infty, \kappa_{jk1} = 0 < \kappa_{jk2} < \cdots < \kappa_{jkM_{jk}} = \infty). \tag{18}$$

In Equation 15, the item threshold (i.e., b_{jk}) and discrimination (i.e., a_{jk}) parameters have a truncated bivariate normal prior with parameters $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$. Notice the $a_{jk} > 0$ restriction (Albert, 1992) assumes θ_{ik} is positively related to item

responses (i.e., higher achieving students are more likely to earn a better item score). Equation 16 is a Beta prior for the 3PNO item guessing parameters (i.e., c_{jk}). Equation 17 indicates that ONO thresholds have a product prior over items, and, similar to Albert and Chib (1993) and Cowles (1996), Equation 18 is an improper uniform prior with threshold monotonicity restrictions for each item.

Multivariate Latent Regression With a Generalized Laplace Prior

This subsection develops a multivariate version of the Bayesian Lasso. The parameter full conditional distributions are reported in Appendix B, available in the online version of the article. Figure 1 includes the directed acyclic graph for the regression portion of the model. For the discussion below, assume the columns of $\mathbf{X}_g$ are centered and orthonormalized (not standardized) within groups. That is, $\mathbf{X}'_g\mathbf{1}_N = \mathbf{0}_{V_g}$ and $\mathbf{X}'_g\mathbf{X}_g = \mathbf{I}_{V_g}$ where $\mathbf{0}_{V_g}$ is a V_g dimensional vector of zeros and $\mathbf{I}_{V_g}$ is a V_g dimensional identity matrix. Note that $\mathbf{X}_g$ is orthonormalized to simplify computations, and Yuan and Lin (2006) discuss how orthonormalizing offers a natural approach for penalizing groups of variables. Furthermore, partition the $K \times V$ matrix of regression coefficients $\mathbf{\Gamma}'$ by letting $\mathbf{\Gamma}' = (\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_G)$ where $\mathbf{\Gamma}_g$ is $K \times V_g$. For instance, if $\mathbf{X}_g$ includes race/ethnicity dummy variables, $\mathbf{\Gamma}_g$ includes coefficients that quantify the subgroup differences across the K content areas.

Bayesian model formulation. The developed Bayesian formulation follows:

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})' | \mathbf{\Gamma}, \mathbf{\Sigma} \sim \text{indep. } \mathcal{N}_K(\mathbf{\Gamma}'\mathbf{x}_i, \mathbf{\Sigma}), \tag{19}$$

$$\text{vec}(\mathbf{\Gamma}_g) | \mathbf{\Sigma}, \tau_g^2 \sim \text{indep.} \mathcal{N}_{KV_g}(\mathbf{0}_{KV_g}, \tau_g^2 \mathbf{I}_{V_g} \otimes \mathbf{\Sigma}), \tag{20}$$

$$p(\tau_g^2 | \lambda^2) \propto (\tau_g^2)^{\frac{KV_g - 1}{2}} \exp\left(-\frac{V_g \lambda^2}{2} \tau_g^2 \right), \tag{21}$$

$$p(\mathbf{\Sigma} | \mathbf{\Sigma}_0, \nu_0) \propto |\mathbf{\Sigma}|^{-\frac{\nu_0 + K + 1}{2}} \exp\left\{ -\frac{1}{2} tr(\mathbf{\Sigma}_0 \mathbf{\Sigma}^{-1}) \right\}, \tag{22}$$

$$p(\lambda^2 | \delta_0) = \frac{\delta_0}{2} \exp\left(-\frac{\delta_0}{2} \lambda^2 \right), \tag{23}$$

where $\otimes$ is a Kronecker product, vec is a column-wise vectorization, and $\mathbf{0}_{KV_g}$ is a KV_g vector of zeros. First, note that Equation 19 is the same as the multivariate regression model employed by NAEP and other testing programs and that the prior for $\mathbf{\Sigma}$ is an inverse Wishart.

Second, the innovation of the developed model resides with the priors in Equations 20, 21, and 23 involving the scale mixture parameters $\tau_1^2, \ldots, \tau_G^2$ and the Lasso parameter λ^2. That is, Equation 20 shows that the prior for $\text{vec}(\mathbf{\Gamma}_g)$ is an MVN scale mixture involving τ_g^2. In contrast, existing methods that use an

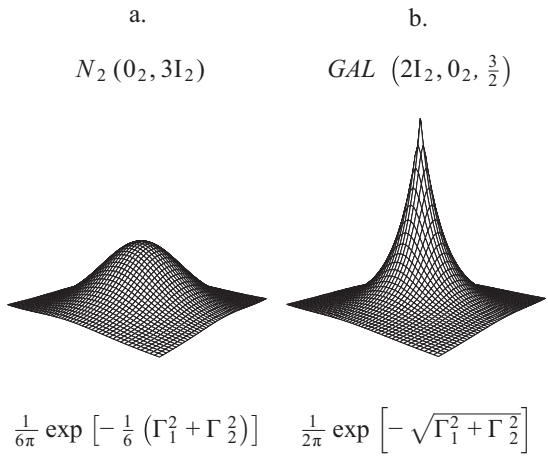

FIGURE 2. *Hypothetical prior densities for Γ_1 and Γ_2. Panel a includes a bivariate normal prior and Panel b includes a generalized asymmetric Laplace prior. Note that a $\mathcal{N}_2(\mathbf{0}_2, 3\boldsymbol{I}_2)$ density where $\mathbf{0}_2$ is a two-dimensional vector of zeros was plotted so the normal and generalized asymmetric Laplace (GAL) distributions have the same marginal variances. For the GAL prior we propose, this would correspond to* $\mathrm{K} = 2$, $\mathrm{V_g} = 1$, $\boldsymbol{\Sigma} = I_2$, *and* $\lambda^2 = 1$.

MVN prior set $\tau_g^2 = 1$ and omit Equations 21 and 23. Introducing τ_g^2 is beneficial, given that smaller values of τ_g^2 in the prior for $\mathrm{vec}(\boldsymbol{\Gamma}_g)$ decreases the variance of coefficients and shrinks coefficients toward the prior mean of zero. Equation 21 indicates $\tau_g^2|\lambda^2 \sim \mathrm{gamma}\left((KV_g+1)/2, V_g\lambda^2/2\right)$ and derivations in Appendix C (available in the online version of the article) demonstrate that the prior for τ_g^2 implies the conditional prior of $\mathrm{vec}(\boldsymbol{\Gamma}_g)|\boldsymbol{\Sigma}, \lambda^2$ is proportional to a multivariate GAL distribution (Kozubowski et al., 2013) defined as

$$p\left(\mathrm{vec}(\boldsymbol{\Gamma}_g)|\boldsymbol{\Sigma}, \lambda^2\right) \propto \exp\left\{-[V_g\lambda^2 tr(\boldsymbol{\Gamma}'_g\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}_g)]^{\frac{1}{2}}\right\}. \tag{24}$$

Figure 2 plots MVN and GAL priors for $\mathrm{vec}(\boldsymbol{\Gamma}_g)$ in the case where $K = 2$, $V_g = 1$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, and $\lambda^2 = 1$. Panel *a* plots a bivariate normal density, $\mathcal{N}_2(\mathbf{0}_2, 3\mathbf{I}_2)$, and panel *b* plots a $\mathrm{GAL}\left(2\mathbf{I}_2, \mathbf{0}_2, \frac{3}{2}\right)$ density (see Appendix in the online version of the article). Note that a bivariate normal with marginal variances of 3 was plotted to match the scale of the GAL distribution in panel *b*. Figure 2 shows that the GAL prior is more peaked at the origin than the bivariate normal density and has heavier tails. The GAL and MVN priors both shrink coefficients to zero, but the heavier tails of the GAL prior give greater weight

to larger coefficients than the MVN prior. One consequence is that the GAL prior can be expected to impose a Bayesian Lasso effect where larger regression coefficients are shrunken less toward zero than the MVN prior thereby improving estimation of significant coefficients.

In Equation 21, each τ_g^2 is dependent upon the Lasso parameter λ^2. The prior implies that $E(\tau_g^2|\lambda^2) = (KV_g + 1)/(V_g\lambda^2)$ so that larger values of λ^2 correspond with smaller values for τ_g^2 and greater shrinkage for $\text{vec}(\boldsymbol{\Gamma}_g)$. Following Park and Casella (2008), an exponential prior with rate parameter $\delta_0/2$ is included for λ^2.

Parameter normalization. It is important to discuss parameter normalization, given that the model does not automatically normalize the distributions of the $\boldsymbol{\theta}_i$. On average, however, they are centered: The expected value of the average $\boldsymbol{\theta}_i$ in the posterior distribution is

$$E\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\theta}_i|\mathbf{y}_1,\ldots,\mathbf{y}_N\right) = E(\boldsymbol{\Gamma}'|\mathbf{y}_1,\ldots,\mathbf{y}_N)\sum_{i=1}^{N}\frac{\mathbf{x}_i}{N}, \tag{25}$$

which is zero because the columns of $\mathbf{X}$ are mean centered. Note that fixing the location of $\boldsymbol{\theta}_i$ ensures that the item threshold parameters b_{jk} are uniquely identified. Second, we employ transformations similar to the parameter expansion approach of Lawrence, Bingham, Liu, and Nair (2008): $\boldsymbol{\theta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ have transformed versions $\widetilde{\boldsymbol{\theta}}$, $\widetilde{\boldsymbol{\Gamma}}$, and $\widetilde{\boldsymbol{\Sigma}}$ that correspond to unit error variances in the latent variable regression. Let $\boldsymbol{\Delta}^2 = \text{diag}(\boldsymbol{\Sigma})$ and define transformed parameters as $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}\boldsymbol{\Delta}^{-1}$, $\widetilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}$, and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}$. The item discrimination parameter matrix is defined as $\mathbf{A} = \bigoplus_{k=1}^{K}\mathbf{a}_k$ (where $\bigoplus$ denotes a direct sum and $\mathbf{a}_k$ is the vector of discrimination parameters for items from area k) and can be transformed to a common metric as $\widetilde{\mathbf{A}} = \mathbf{A}\boldsymbol{\Delta}$.

Rescaling regression coefficients to the original metric. Recall that $\mathbf{X}_g$ was assumed to be orthonormalized within group g. Researchers may be interested in rescaling the coefficients to the original metric for the predictor variables. Consider the QR decomposition $\mathbf{X}_g^* = \mathbf{X}_g\mathbf{R}_g$, where $\mathbf{X}_g^*$ contains the original predictor variables in group g. The rescaled, transformed regression coefficients for group g are computed as $\widetilde{\boldsymbol{\Gamma}}_g^* = \widetilde{\boldsymbol{\Gamma}}_g(\mathbf{R}_g^{-1})'$.

Assessing statistical significance. This subsection provides details on how researchers can summarize the simulated values from the parameter posterior distributions to assess statistical significance of groups of variables. One option is to transform $\widetilde{\boldsymbol{\Gamma}}$ to the original metric for $\mathbf{X}$ (i.e., $\widetilde{\boldsymbol{\Gamma}}_1^*,\ldots,\widetilde{\boldsymbol{\Gamma}}_G^*$) and construct Bayesian credible intervals to test V coefficients for the K content areas. However, constructing intervals for VK coefficients may be less preferred as V and K increase.

A second option employed below is to assess the multivariate statistical significance of groups of variables by comparing the distance of the origin from the posterior distribution for each $\text{vec}(\widetilde{\mathbf{\Gamma}}_g)$. Let $t = 1, \ldots, T$ index MCMC iterations and let $\text{vec}(\widetilde{\mathbf{\Gamma}}_g)_t$ denote the coefficients for group g at iteration t. Recall that $\text{vec}(\widetilde{\mathbf{\Gamma}}_g)_t$ is a KV_g dimensional vector. The multivariate Mahalanobis distance between $\text{vec}(\widetilde{\mathbf{\Gamma}}_g)_t$ for each MCMC iteration with the respective centroid of the estimated posterior distribution is

$$D_{gt} = \left(\left[\text{vec}(\widetilde{\mathbf{\Gamma}}_g)_t - \overline{\text{vec}(\widetilde{\mathbf{\Gamma}}_g)} \right]' \widehat{\mathbf{\Sigma}}_{\Gamma g}^{-1} \left[\text{vec}(\widetilde{\mathbf{\Gamma}}_g)_t - \overline{\text{vec}(\widetilde{\mathbf{\Gamma}}_g)} \right] \right)^{\frac{1}{2}}, \tag{26}$$

where $\overline{\text{vec}(\widetilde{\mathbf{\Gamma}}_g)}$ denotes the average and $\widehat{\mathbf{\Sigma}}_{\Gamma g}$ the sample variance–covariance matrix computed from the posterior draws. Note that the Bayesian approach yields a distribution for D_{gt} and the distance between the group centroid and origin is used to assess statistical difference from zero as

$$D_{g0} = \left(\overline{\text{vec}(\widetilde{\mathbf{\Gamma}}_g)}' \widehat{\mathbf{\Sigma}}_{\Gamma g}^{-1} \overline{\text{vec}(\widetilde{\mathbf{\Gamma}}_g)} \right)^{\frac{1}{2}}. \tag{27}$$

Accordingly, $\hat{p}_g = \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}(D_{g0} < D_{gt})$ provides a pseudo p value for assessing whether the coefficients for group g differ from zero. D_{g0} will be larger and $\hat{p}_g$ will be smaller for groups with coefficients that differ from $\mathbf{0}_{KV_g}$.

Neither the MVN nor GAL by themselves provide variable selection capabilities. Thus, we use $\hat{p}_g$ to perform a kind of thresholding to select variables. For example, we select all variable groups with $\hat{p}_g < 0.05$.

It is important to note that Mahalanobis distance provides an adequate measure of distance for ellipsoidal multivariate distributions. Plots of $\text{vec}(\widetilde{\mathbf{\Gamma}}_g)$ that were not included in the article provided evidence of approximately ellipsoidal distributions for all 148 variable groups, which suggested that D_{g0} is an adequate measure of distance from the origin in the posterior.

Monte Carlo Simulation

This section presents Monte Carlo evidence to compare the GAL and MVN priors in terms of estimation and model selection accuracy. Results from two Monte Carlo simulation studies are presented to consider the relative performance of the GAL and MVN for different levels of variable group sparsity (e.g., 80% variable group sparsity for Simulation #1 and 40% variable group sparsity for Simulation #2) and whether the predictor relationships differed by latent dimension.

Simulation Study #1

Overview. The simulation expands upon the univariate design of Yuan and Lin (2006). First, a predictor matrix $\mathbf{X}$ with $V = 30$ columns was generated by

sampling latent variables Z_g for $g = 1, \ldots, G = 15$ from a standardized MVN distribution with correlations between variable g and g' equal to $0.5^{|g-g'|}$. Predictors were created with $V_g = 2$ by trichotomizing Z_g as

$$X_{1g} = \begin{cases} 1 & Z_g > \Phi^{-1}\left(\frac{2}{3}\right) \\ 0 & \text{otherwise} \end{cases}, \; X_{2g} = \begin{cases} 1 & Z_g < \Phi^{-1}\left(\frac{1}{3}\right) \\ 0 & \text{otherwise} \end{cases}, \tag{28}$$

so that middle values of Z_g are the reference category. The $K = 3$ latent dependent variables were generated as

$$\theta_k = 1.8X_{11} - 1.2X_{21} + X_{13} + 0.5X_{23} + X_{15} + X_{25} + \varepsilon_k. \tag{29}$$

Note that the data generating model for Simulation Study #1 represented a sparse condition where 80% of the rows are zero. The residual variance–covariance matrix $\boldsymbol{\Sigma}$ included ones on the diagonal and correlations equal to .5. The signal to noise ratio ψ was manipulated with values of 0.5, 1.0, and 1.8. The simulation study also considered sample sizes of $N = 250$, 500, 1,000, and 2,500.

Second, the θ_k were normalized to a variance of one to generate 10 items per dimension for a total of $J = 30$ items. The items were generated from the 2PNO item response model with item parameters sampled each replication as $a \sim \mathcal{N}(1.1, 0.05)\mathcal{I}(a > 0)$ and $b \sim \text{uniform}(-1.5, 1.5)$.

Third, researchers need guidance for selecting a value for δ_0 in the exponential prior for λ^2 in Equation 23. The simulation considered $\delta_0 = 10^{-1}$ and 10^{-4}, which implies prior expected values for λ^2 of 20 and 20,000, respectively. Note that $\delta_0 = 10^{-4}$ is associated with a flatter prior for λ^2 than 10^{-1} and the simulation study accordingly assesses the influence of employing a more or less informative prior for λ^2 on parameter estimation.

The simulation employed a crossed design with four values for N, three values for ψ, two values for δ_0, and 100 replications per condition for a total of 2,400 data sets. The MCMC GAL and MVN algorithms were implemented with 25,000 iterations and the first 12,500 iterations were discarded as burn-in. The replications with $N = 2{,}500$ required approximately 70 min to complete both the GAL and MVN models using a cluster with Intel HP X5650 2.66Ghz 6C processors. The prior parameters for item parameters were $\boldsymbol{\mu}'_\xi = (0, 0)$ for the mean and $\boldsymbol{\Sigma}_\xi = \text{diag}(4, 16)$ for the variance to yield a more vague prior for $\boldsymbol{\xi}_{jk}$. The prior for the residual variance–covariance matrix for the structural model was defined as an inverse Wishart with parameters $\boldsymbol{\Sigma}_0$ and "prior sample size" ν_0 (Hoff, 2009). Choosing $\boldsymbol{\Sigma}_0 = \mathbf{I}_K$ is reasonable, given that $\boldsymbol{\Sigma}$ is transformed to have ones along the diagonal. Additionally, selecting $\nu_0 = K + 2$ is the smallest integer-valued prior sample size that can be specified, so that the prior expected value of $\boldsymbol{\Sigma}$ exists.

The GAL and MVN approaches were evaluated based upon model selection accuracy and the bias and variance for $\boldsymbol{\Gamma}$. That is, the GAL was hypothesized to

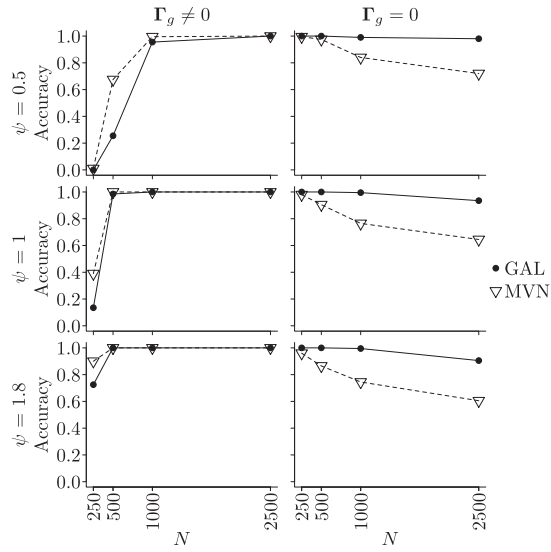

FIGURE 3. *Model selection accuracy rates for Simulation Study #1 by nonzero and zero regression coefficients by sample size and signal-to-noise ratio* ψ. *Accuracy is computed as the proportion of times* $\hat{\mathrm{p}}_{\mathrm{g}}$ *correctly selected all nonzero and zero regression coefficients, using the threshold of 0.05. The proportions were computed over both* δ_0 *conditions, so each point is based upon 200 replications.*

provide more accurate model selection. The simulation recorded the number of times using $\hat{p}_g$ with a 0.05 rejection level correctly selected all variable groups with nonzero and zero regression coefficients in $\boldsymbol{\Gamma}$.

Results. Figure 3 plots the accuracy of using $\hat{p}_g$ to select the nonzero and zero variable group regression coefficients for the GAL and MVN priors by N and ψ. Note the manipulated values of δ_0 did not impact model selection, so the results were aggregated over δ_0 so each point is based upon 200 replications. The first column of plots in Figure 3 reports the proportion of times $\hat{p}_g$ correctly indicated $\boldsymbol{\Gamma}_g \neq \mathbf{0}$ for $g = 1, 3, 5$. The results suggest the MVN prior was more powerful for smaller N and ψ, but the difference declined as both ψ and N increased. The second column of plots in Figure 3 includes the proportion of times at least one $\hat{p}_g$ classified a zero regression coefficient group significant for $g = 2, 4, 6, \ldots, 15$. The GAL and MVN priors correctly identified the $\boldsymbol{\Gamma}_g = \mathbf{0}$ variable groups when $N = 250$, but the model selection accuracy for the MVN declined as both N and ψ increased. In contrast, the GAL prior correctly identified $\boldsymbol{\Gamma}_g = 0$ with over 90% accuracy for all combinations of ψ and N. For instance, for $\psi = 1.8$ and $N =$ 2,500, the GAL and MVN correctly identified all of the $\boldsymbol{\Gamma}_g = 0$ variable groups in 91% and 61% of the replications.

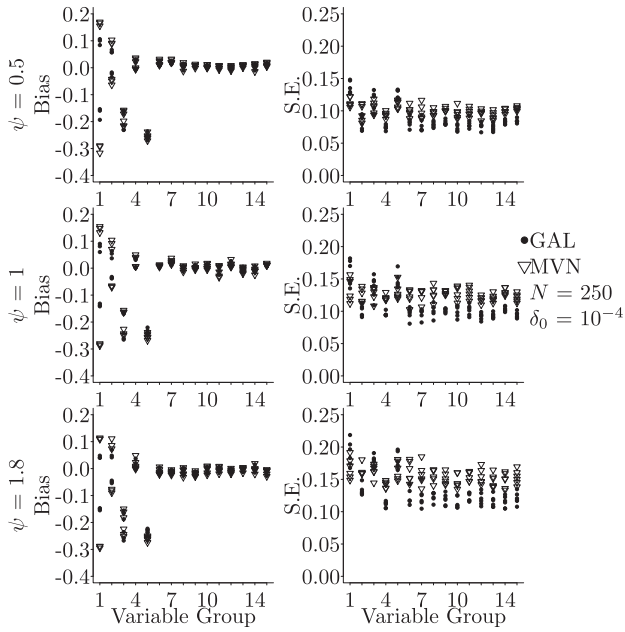

FIGURE 4. *Bias and standard errors of generalized asymmetric Laplace and multivariate normal estimates for Simulation Study #1 by* ψ *and variable group for* $\mathrm{N} = 250$ *and* $\delta_0 = 10^{-4}$. *Bias and standard errors are reported for all* $\mathrm{KV} = 90$ *regression coefficients and were computed from 100 replications.*

Figures 4 and 5 report the bias and standard errors for the $KV = 90$ GAL and MVN regression coefficients by ψ for $N = 250$ and 2,500, respectively, and $\delta_0 = 10^{-4}$. Note that eight figures for all values of N and δ_0 are provided as Supplemental Files. Figure 4 provides evidence that the GAL prior yields less biased estimates for variable Groups 1, 3, and 5 (i.e., the groups where $\boldsymbol{\Gamma}_g \neq \mathbf{0}$). For example, the absolute bias for elements of the largest regression coefficients $\boldsymbol{\Gamma}_1$ exceeded 0.2 for the MVN in contrast to values between 0.1 and 0.2 for the GAL prior. The bias for insignificant variable groups was near zero for both the GAL and MVN with one notable exception. That is, the MVN was more biased than the GAL for variable group $g = 2$. Recall that the predictors were generated to correlate most with adjacent variables, so $\mathbf{X}_2$ was most related to two significant groups $\mathbf{X}_1$ and $\mathbf{X}_3$. The simulation results provide evidence that the GAL performs better than the MVN in dealing with correlated groups of variables. The second column of plots in Figure 4 provides evidence of differences in variability between the GAL and MVN priors. Specifically, the variability of the GAL estimates for the $\boldsymbol{\Gamma}_g = \mathbf{0}$ coefficients were systematically smaller. In contrast, the GAL estimates were more variable for the significant variable groups $g = 1, 3, 5$.

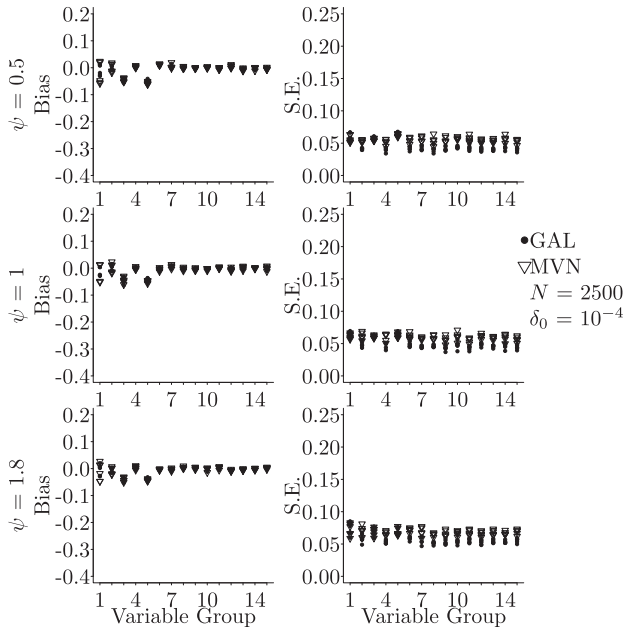

FIGURE 5. *Bias and standard errors of generalized asymmetric Laplace and multivariate normal estimates for Simulation Study #1 by* ψ *and variable group for* N = *2,500 and* $\delta_0 = 10^{-4}$. *Bias and standard errors are reported for all* KV = *90 regression coefficients and were computed from 100 replications.*

Figure 5 reports GAL and MVN bias and standard errors for $N = 2{,}500$. As noted above, the GAL and MVN models are both consistent, and Figure 5 shows the bias declines for both priors for larger N. The greatest difference was for $\mathbf{\Gamma}_1$; however, the GAL and MVN were more similar for $N = 2{,}500$ than for $N = 250$. Figure 5 provides evidence that the standard errors for the GAL are smaller than the MVN for the $\mathbf{\Gamma}_g = \mathbf{0}$ variable groups. Unlike Figure 4, the GAL and MVN standard errors are similar for significant variable groups.

Simulation Study #2

Overview. The second simulation study changed three features of Simulation Study #1. First, the prior parameter was set as $\delta_0 = 10^{-4}$ for 200 replications. Second, the matrix of regression coefficients $\mathbf{\Gamma}$ was modified to have 40% of its rows equal to zero. Third, Simulation Study #2 varied the significant predictors variables by latent trait. Specifically, the latent traits for $k = 1, 2, 3$ related to the variable groups as

$$\theta_k = 1.8X_{1,k} - 1.2X_{2,k} + X_{1,k+4} + 0.5X_{2,k+4} + X_{1,k+8} + X_{2,k+8} + \varepsilon_k. \qquad (30)$$

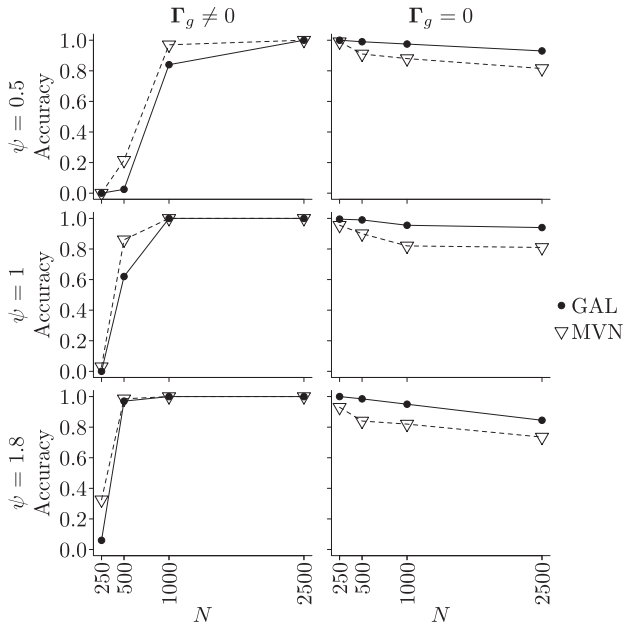

FIGURE 6. *Model selection accuracy rates for Simulation Study #2 by nonzero and zero regression coefficients by sample size and signal-to-noise ratio* ψ. *Accuracy is computed as the proportion of times* $\hat{\mathrm{p}}_{\mathrm{g}}$ *correctly selected all nonzero and zero regression coefficients, using the threshold of 0.05. The proportions were computed over both* δ_0 *conditions, so each point is based upon 200 replications.*

Results. Figure 6 plots the accuracy of using $\hat{p}_g$ to select the nonzero and zero variable group regression coefficients for the GAL and MVN priors by N and ψ. The results in the first column for $\mathbf{\Gamma} \neq 0$ shows the MVN prior was more powerful for smaller $N \leq 500$ and $\psi = 0.5$, but the difference declined as both ψ and N increased. The second column in Figure 6 reports results concerning the classification of zero regression coefficients. The GAL and MVN priors correctly identified the $\mathbf{\Gamma}_g = \mathbf{0}$ variable groups when $N = 250$ and $\psi \leq 1.0$, but the relative model selection accuracy for the MVN declined as both N and ψ increased. In contrast, the GAL prior outperformed the MVN across N and ψ.

Figure 7 reports bias and standard errors of the GAL and MVN estimators for the $N = 250$ and $\psi = 0.5, 0.1, 1.8$. The GAL demonstrated less bias than the MVN for the variables groups with the largest magnitude coefficients (i.e., the largest coefficients correspond to $g = 1$ for θ_1, $g = 2$ for θ_2, and $g = 3$ for θ_3). Figure 7 also provides evidence that smaller nonzero coefficients are shrunken more under the GAL than the MVN, given that the bias for variable groups $g = 5$, 6, 7, 9, 10, 11 is more negative under the GAL.

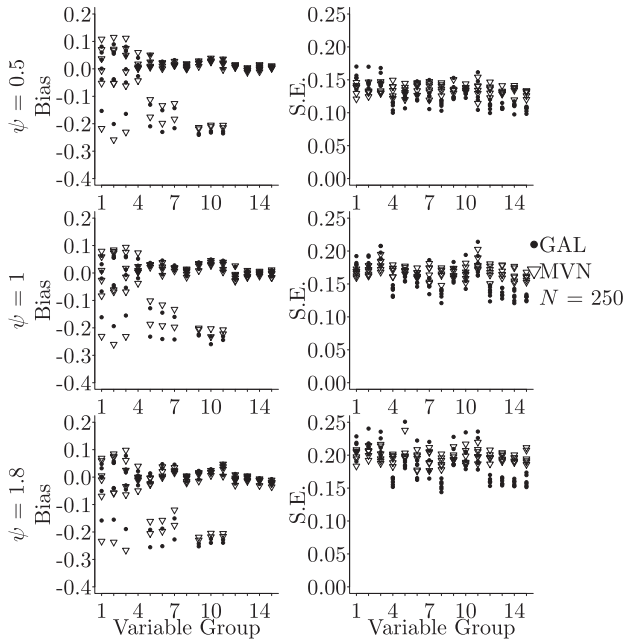

FIGURE 7. *Bias and standard errors of generalized asymmetric Laplace and multivariate normal estimates for Simulation Study #2 by* ψ *and variable group for* $N = 250$ *and* $\delta_0 = 10^{-4}$. *Bias and standard errors are reported for all* $KV = 90$ *regression coefficients and were computed from 100 replications.*

The results in Figure 7 provide insight regarding the difference in model selection performance between the GAL and MVN in Figure 6 for $N = 250$. In particular, the smaller coefficients are biased more toward zero under the GAL for smaller sample sizes. One implication is that the GAL may be less likely to identify smaller coefficients when $N < 1{,}000$.

Figure 8 reports bias and standard errors of the GAL and MVN estimators for the $N = 2{,}500$ and $\psi = 0.5, 1.0, 1.8$. The pattern of bias is similar to the $N = 250$ case with the exception that the degree of bias is generally small. Similar to Figures 4 and 5, Figures 7 and 8 show that standard errors for the zero coefficients are smaller for the GAL than the MVN prior.

Application to the NAEP Data

This section reports an application of the developed multivariate latent regression model to the 2011 NAEP mathematics data. The first section discusses the data set and MCMC implementation and the second section reports results.

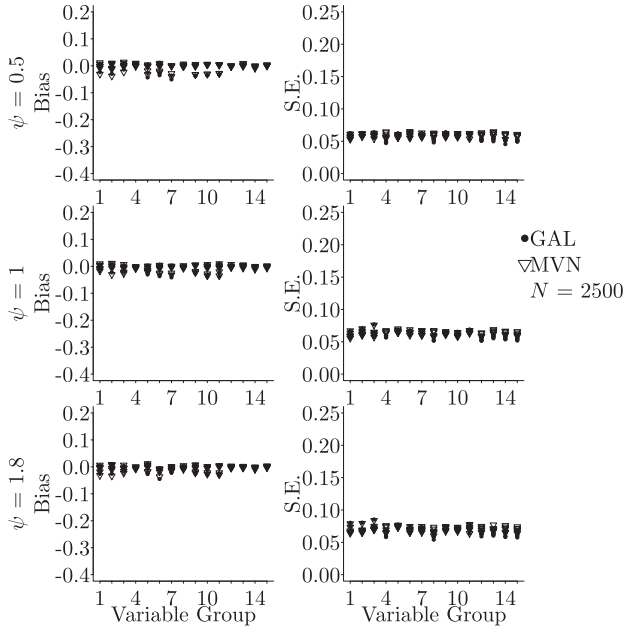

FIGURE 8. *Bias and standard errors of generalized asymmetric Laplace and multivariate normal estimates for Simulation Study #2 by* ψ *and variable group for* N = *2,500 and* $\delta_0 = 10^{-4}$. *Bias and standard errors are reported for all* KV = *90 regression coefficients and were computed from 100 replications.*

Sample, Variables, Prior Parameters

In 2011, NAEP administered $J = 155$ items to 175,200 8th-grade students to assess mathematics achievement in $K = 5$ subject areas: (1) algebra ($J_1 = 49$); (2) data analysis, statistics, and probability ($J_2 = 23$); (3) geometry ($J_3 = 30$); (4) measurement ($J_4 = 26$); and (5) number properties and operations ($J_5 = 27$). Furthermore, the regression model included $G = 148$ groups with a total of $V = 261$ variables and $KV = 1{,}305$ regression coefficients.

Some students had missing cognitive responses. "Omitted" responses were scored as incorrect (e.g., NAEP documentation defines "Omitted" as questions without a response, but responses on adjacent items) and "Not Reached" were treated as missing at random, given that item booklets were not sorted from least to most difficult items. Many of the background variables were considered as factors (e.g., four-category Likert-type responses may be modeled as a group with three dummy variables) and missing responses were included as another factor level. Missing values on continuous background variables, such as years of teaching, were imputed with the mean.

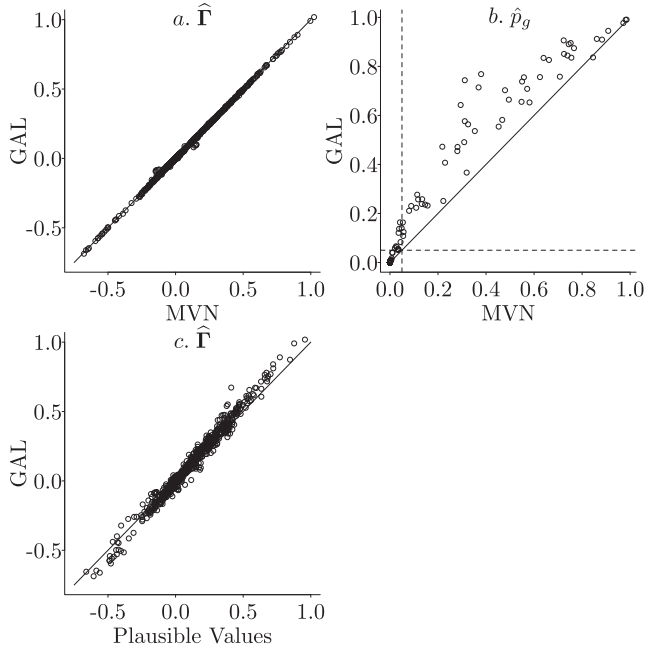

FIGURE 9. *Summary of application to 2011 National Assessment of Educational Progress mathematics data. Panel a plots the estimated regression coefficients for the generalized asymmetric Laplace (GAL) and multivariate normal (MVN) models. Panel b reports the pseudo* p *values* $\hat{\text{p}}_{\text{g}}$ *for the GAL and MVN models. Panel c plots the estimated coefficients for the GAL and plausible values regression using the AM software. 45° reference lines are included in all three panels.*

MCMC was employed, and 50,000 iterations were run with 20,000 treated as burn-in. Similar to the simulation study, the prior parameters were defined as $\boldsymbol{\mu}_\xi = \mathbf{0}_2$, $\boldsymbol{\Sigma}_\xi = \mathbf{I}_2$, $\alpha_c = \beta_c = 1$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_K$, $\delta_0 = 10^{-4}$, and $\nu_0 = K + 2$. A standard deviation equal to $0.175/M_{jk}$ was employed for the proposal distribution for the Metropolis–Hasting sampler for item category thresholds (e.g., see Cowles, 1996), so that the proposal standard deviations differed based upon the number of item categories. The proposal standard deviation yielded threshold acceptance rates between 25% and 50%.

Results

Figure 9 provides a comparison among the GAL, MVN, and AM software results. Panel *a* shows that the GAL and MVN estimates were nearly identical, which is expected based upon the Monte Carlo results given that N was large. Panel *b* of Figure 9 plots $\hat{p}_g$ for the GAL and MVN and provides evidence that the

MVN $\hat{p}_g$ were generally smaller. In fact, the MVN selected 95 of the 148 variable groups whereas the GAL selected 83. Panel *b* shows there were no instances where the GAL selected groups that were not already selected by the MVN. Recall that the Monte Carlo study found evidence that the GAL was more parsimonious in terms of model selection, and the application to the NAEP data demonstrates a similar pattern.

Panel *c* compares estimated regression coefficients between the GAL and PV regressions using the AM software. Note the PV coefficients were divided by the corresponding estimated error standard deviation, so both GAL and AM coefficients can be interpreted with error variances equal to one. The results in Panel *c* provide evidence of more differences between the GAL and PV coefficients than the GAL and MVN coefficients in Panel *a*. The differences between the GAL and PV in Panel *c* of Figure 9 may also translate to practical differences for researchers and policymakers who are interested in subgroup differences. That is, the results in Panel *c* provide evidence that PV regression coefficients tended to be smaller in magnitude than the GAL. Table 1 provides an example of differences between the GAL and PV coefficients by reporting race/ethnicity regression coefficients for all five content areas (note that multiracial students were the reference group and the results are unweighted). Table 1 provides evidence that GAL and PV differ in terms of some achievement gap interpretations. For instance, the coefficients for African American students were noticeably smaller for the PV versus the GAL on the data analysis, statistics, and probability (i.e., -0.383 vs. -0.515) and number properties and operations (i.e., -0.478 vs. -0.595) dimensions. The coefficients for the GAL were generally smaller for the American Indian/Alaska Native comparison than PV, but the differences could be partially explained by larger standard errors due to a smaller sample size.

Discussion

This section summarizes the contributions of the article to existing research and current statistical practice in large-scale testing programs and offers comment regarding future research efforts. First, this study contributes to the literature by introducing a novel Bayesian procedure that provides approximate model selection for multivariate regression models as found in large-scale assessments. In fact, results from two simulation studies provide evidence the GAL yields more accurate model selection in larger samples and reduced bias for significant variable groups. The difference in model selection accuracy between the GAL and MVN is likely attributed to the improved precision in estimating zero coefficients. That is, standard errors for the GAL were systematically smaller than the MVN for insignificant variable groups, and there was accordingly a smaller chance that $\hat{p}_g < 0.05$. Furthermore, the GAL also outperformed the MVN for estimating coefficients of groups that correlated with significant variable groups,

TABLE 1.
Race-Based Achievement Gaps Using the GAL Model and Plausible Values (PV)

		GAL		PV	
Race	Content Area	EST	*SE*	EST	*SE*
African Am.	1	−0.498	0.027	−0.440	0.031
	2	−0.515	0.034	−0.383	0.030
	3	−0.559	0.031	−0.471	0.025
	4	−0.595	0.030	−0.478	0.035
	5	−0.573	0.029	−0.489	0.025
Am. Indian/Alaska Native	1	−0.274	0.037	−0.350	0.048
	2	−0.399	0.045	−0.434	0.052
	3	−0.170	0.041	−0.243	0.048
	4	−0.259	0.040	−0.307	0.064
	5	−0.322	0.040	−0.403	0.049
Asian	1	0.485	0.030	0.451	0.039
	2	0.357	0.038	0.351	0.028
	3	0.364	0.034	0.354	0.035
	4	0.422	0.034	0.421	0.038
	5	0.377	0.034	0.326	0.030
Hispanic of any race	1	−0.080	0.028	−0.051	0.035
	2	−0.078	0.034	−0.036	0.023
	3	−0.024	0.031	0.000	0.030
	4	−0.040	0.030	0.017	0.045
	5	−0.114	0.030	−0.101	0.033
Native/Pacific Islander	1	−0.123	0.051	−0.150	0.062
	2	−0.166	0.062	−0.096	0.067
	3	−0.129	0.059	−0.113	0.047
	4	−0.175	0.057	−0.156	0.054
	5	−0.244	0.055	−0.186	0.053
White	1	0.084	0.026	0.096	0.029
	2	0.066	0.031	0.088	0.026
	3	0.082	0.029	0.093	0.029
	4	0.086	0.028	0.120	0.032
	5	0.032	0.028	0.032	0.030

Note. Results are unweighted. EST = estimated posterior mean; *SD* = estimated posterior standard deviation; Am. = American; GAL = generalized asymmetric Laplace. The content areas are 1 = algebra; 2 = data analysis, statistics, and probability; 3 = geometry; 4 = measurement; and 5 = number properties and operations.

which is important given that large-scale testing programs include many correlated predictors. Lastly, the Monte Carlo results suggest researchers may prefer using the GAL in smaller samples given the reduced bias in estimating $\mathbf{\Gamma}$.

Second, the improved approximate model selection of the GAL versus MVN has implications for test developers and large-scale testing programs. NAEP is

designed to broadly sample content and scores are reported for groups of students rather than for individual students, which is different from district and state standardized testing programs. Instead, responses to hundreds of background questions are used to predict the performance for groups of students. The ability of NAEP to satisfy its mission of providing data about education relies upon the quality of predictors used in the conditioning model. The developed model could be used to support the redesign of background questionnaires. That is, test developers can estimate a high-dimensional regression with a GAL prior to assess which background questions uniquely contribute to the prediction of student achievement. For instance, the application to the 2011 NAEP mathematics data suggests that 83 of the 148 variables had a statistically significant relationship with achievement. Consequently, it is possible that students, teachers, and school administrators dedicate time to answer at least 65 additional survey questions that provide no unique contribution to predictions. The GAL prior could be used to decide which background questions to retain, modify, or delete for subsequent data collections to optimize the time students, teachers, and school administrators dedicate to completing surveys.

Third, there were noticeable differences in model estimates for the application to the 2011 NAEP mathematics data between the GAL and PV using the AM software. An example was provided in Table 1 showing that the size of the achievement gap for African American students was larger based upon the GAL than PV. Consequently, policymakers could make different conclusions about the status of the achievement gap based upon their choice of data analytic strategy. One additional difference relates to model selection. That is, the GAL can use $\hat{p}_g$ to select variable groups. In contrast, the AM software utilizes univariate PV regressions and does not offer researchers tools to select groups of variables in a multivariate regression.

Fourth, this article contributes to the statistics literature by extending the Bayesian Lasso to multivariate contexts. The multivariate Bayesian Lasso provides a natural approach for modeling multivariate latent variables and offers a computationally efficient Gibbs sampler. Although the central focus of this article was on NAEP, the developed model is directly relevant to other large-scale international testing programs and other research domains involving either observed or latent multivariate dependent variables.

There are several directions for future research. First, Park and Casella (2008) showed how the univariate Bayesian Lasso formulation ensures a unimodal posterior distribution. The results of Park and Casella should directly extend to the multivariate Bayesian group Lasso developed in this article in cases where the vectors of dependent variables are observed. There was no evidence of multimodal posterior distributions for the application of this study and future research is needed to prove the proposed formulation ensures a unimodal posterior distribution for latent dependent variables. Second, as noted by an anonymous

reviewer, the developed model assumed predictor variables were missing at random. Future research should consider methods for performing model selection in the presence of nonignorable missing data. Third, the application of the developed model did not account for the sampling design of NAEP. Future research should examine approaches for incorporating sampling weights into multivariate Bayesian regression models.

In conclusion, the methods developed in this article improve upon standard Bayesian procedures for multivariate latent regression models. Evidence was presented to support the preference for a Laplace versus Gaussian prior when making inferences about model selection in large-scale testing.

Authors' Note

Opinions reflect those of the author(s) and do not necessarily reflect those of the granting agencies.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF Grant DRL-0941014.

References

Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*, 251–269.

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

American Institutes for Research. (2002–2012). *AM statistical software*. Retrieved from http://am.air.org/default.asp

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*, 101–111.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York: Springer-Verlag.

Fox, J. P. (2010). *Bayesian item response modeling*. New York, NY: Springer.

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

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

Gonzalez, E., & Rutkowski, L. (2010). Practical approaches for choosing multiple-matrix sample designs. *IEA-ETS Research Institute Monograph*, *3*, 125–156.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika*, *96*, 835–845.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. Boca Raton, FL: CRC Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Dordrecht: Springer.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*, 95–110.

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

Johnson, M. S., & Jenkins, F. (2004). *A Bayesian hierarchical model for large-scale educational surveys: An application to the national assessment of educational progress*. Princeton, NJ: Educational Testing Service.

Johnson, M. S., & Sinharay, S. (2015). *Does the NAEP model adequately predict the achievement gap?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kozubowski, T. J., Podgórski, K., & Rychlik, I. (2013). Multivariate generalized Laplace distribution and related random fields. *Journal of Multivariate Analysis*, *113*, 59–72.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, *5*, 369–411.

Lawrence, E., Bingham, D., Liu, C., & Nair, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics*, *50*, 182–191.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*, 245–256.

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

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

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions* (Tech. Rep. No. RR-05-27). Princeton, NJ: Educational Testing Service.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, *25*, 351–371.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, *22*, 425–445.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

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

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

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

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 49–67.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*, 293–308.

Authors

STEVEN ANDREW CULPEPPER is an associate professor in the Department of Statistics at the University of Illinois at Urbana-Champaign, Illini Hall, Room 115, 725 S. Wright Street, Champaign, IL 61820; email: sculpepp@illinois.edu. His research interests are statistical modeling in the social sciences, Bayesian models and computation, restricted latent class models, longitudinal cognitive diagnosis, and innovative standardized testing formats.

TREVOR PARK is a clinical assistant professor of statistics at the University of Illinois at Urbana-Champaign, Illini Hall, Room 101, 725 S. Wright Street, Champaign, IL 61820; email: thp2@illinois.edu. His research interests are multivariate exploratory methods, covariance modeling, optimization methods, the Bayesian Lasso, psychometric analysis of testing data, and distance correlation.

Manuscript received October 10, 2015
First revision received February 26, 2016
Second revision received August 18, 2016
Third revision received January 27, 2017
Accepted January 31, 2017