Article

Journal of Educational and Behavioral Statistics
2018, Vol. 43, No. 6, pp. 663–692
DOI: 10.3102/1076998618795124

Flexible Bayesian Models for Inferences From Coarsened, Group-Level Achievement Data

J. R. Lockwood
Katherine E. Castellano
Educational Testing Service

Benjamin R. Shear
University of Colorado Boulder

This article proposes a flexible extension of the Fay–Herriot model for making inferences from coarsened, group-level achievement data, for example, school-level data consisting of numbers of students falling into various ordinal performance categories. The model builds on the heteroskedastic ordered probit (HETOP) framework advocated by Reardon, Shear, Castellano, and Ho by allowing group parameters to be modeled with regressions on group-level covariates, and residuals modeled using the flexible exponential family of distributions recommended by Efron. We demonstrate that the alternative modeling framework, termed the "Fay–Herriot heteroskedastic ordered probit" (FH-HETOP) model, is useful for mitigating some of the challenges with direct maximum likelihood estimators from the HETOP model. We conduct a simulation study to compare the costs and benefits of several methods for using the FH-HETOP model to estimate group parameters and functions of them, including posterior means, constrained Bayes estimators, and the "triple goal" estimators of Shen and Louis. We also provide an application of the FH-HETOP model to math proficiency data from the Early Childhood Longitudinal Study. Code for estimating the FH-HETOP model and conducting supporting calculations is provided in a new package for the R environment.

Keywords: *ordinal data; heteroskedastic ordered probit model; small-area estimation; student achievement*

Increased standardized testing and data archiving by public education systems have provided unprecedented opportunities for studying these systems and their effects on student outcomes. While data housed in state longitudinal data systems often contain test scores for individual students as they progress through grade levels, it is still commonplace for publicly available data on academic achievement to be much less fine-grained. For example, state department of education websites often provide achievement measures for schools and/or districts consisting of the counts, or percentages, of students falling into a small number of

ordinal performance-level categories (e.g., below basic, basic, proficient, and advanced). Such data lose resolution in two ways relative to individual-level student test scores: They are aggregated to a group level, and they are coarsened by collapsing the scores into a small number of categories.

As argued by Reardon, Shear, Castellano, and Ho (2017), while such aggregate, coarsened data are suitable for some purposes, there is often a desire to use these data to make inferences about achievement on a latent continuous scale, where familiar metrics such as standardized effect sizes and intraclass correlation coefficients (ICCs) are available. Reardon, Shear, et al. (2017) apply the heteroskedastic ordered probit (HETOP) model to provide such inferences. The method takes aggregate, coarsened data from $g = 1, \ldots, G$ groups (e.g., schools or districts), such as counts of how many individuals in each group fall into each of K ordinal performance categories, and computes an estimated mean $\widehat{\mu}_g$ and standard deviation (*SD*) $\widehat{\sigma}_g$ of the distribution of latent, continuous achievement across individuals in each group. The framework is useful for synthesizing data across geographic regions that may have different assessment frameworks and/or performance-level definitions. For example, Fahle and Reardon (2018) use the HETOP model to describe patterns in district-level ICCs of math and English-language arts achievement using aggregate, coarsened achievement data from all U.S. public school districts.

Reardon, Shear, et al. (2017) use the observed count data and the HETOP model to compute what we will call "direct estimates" (DE) of the group parameters $\{\mu_g, \sigma_g\}_{g=1}^{G}$. These estimates are obtained by maximum likelihood estimation (MLE) under the assumption that the group parameters are fixed, unknown constants. The term "direct estimates" is borrowed from the small-area estimation literature (e.g., Ghosh & Rao, 1994; Pfefferman, 2002) and indicates that the parameter estimates for a given group are not informed by either the data for groups with similar observable characteristics or the distributional properties of the true parameters across the ensemble of groups.[1]

The DE computed by MLE have several limitations. First, the MLE has relatively restrictive conditions for its existence (Haberman, 1980; McCullagh, 1980). Existence problems begin to arise when there is at least one group with nonzero counts in fewer than three of the K performance-level categories. With $K = 3$ or 4 typical in applications involving achievement tests, and with many groups (some of which may be small), it is likely that the MLE of the ensemble of true group parameters $\{\mu_g, \sigma_g\}_{g=1}^{G}$ does not exist in a given data set. Second, even when it exists, the MLE can have large estimation errors when group sample sizes are small and/or the marginal probability of at least one of the K performance-level categories is small. The estimation errors can include notable negative bias in $\widehat{\sigma}_g$ under these circumstances (Reardon, Shear, Castellano, and Ho, 2017). The estimation errors in the group parameters can lead to noisy and

biased estimates of functions of those parameters such as the ICC, standardized mean differences between pairs of groups, and distribution functions. For example, the empirical distribution function of $\{\widehat{\sigma}_g\}_{g=1}^{G}$ can be an excessively biased estimator of the distribution function of $\{\sigma_g\}_{g=1}^{G}$ due to overdispersion of the DE $\{\widehat{\mu}_g, \widehat{\sigma}_g\}_{g=1}^{G}$. Thus, the DE are not well suited to support some desired inferences about the true parameters, similar to the arguments provided by Mislevy, Beaton, Kaplan, and Sheehan (1992) regarding MLEs of achievement attributes from item response theory (IRT) models.

Additional issues can arise in some settings from the fact that the DE do not use auxiliary information, which may include group-level covariates (e.g., aggregate demographic characteristics of group members) and/or information about distributional properties of the true parameters across the ensemble of groups. DE of group parameters generally will be less accurate than estimates that synthesize the observed count data with auxiliary information using various forms of shrinkage (Efron & Morris, 1973, 1977; Morris, 1983). The lower accuracy of the DE relative to alternatives may not be problematic, depending on the ultimate goals of the analysis, but direct estimation by MLE provides no mechanism for incorporating auxiliary information in cases where using it may be desirable. Further, when an analyst possesses both group-level covariates and the count data, as might be common with data obtained from state department of education websites and merged to sources such as the National Center for Education Statistics (NCES) Common Core of Data on school and district characteristics (https://nces.ed.gov/ccd), using the DE requires a two-stage procedure to study the relationships between covariates of interest and $\{\mu_g, \sigma_g\}_{g=1}^{G}$: The first stage computes DE $\{\widehat{\mu}_g, \widehat{\sigma}_g\}_{g=1}^{G}$ from the count data, and the second stage uses these estimates as outcome variables in regression models on the covariates. It may be more efficient, and would permit more straightforward assessments of uncertainty, to study relationships between group parameters and covariates within the context of a single model.

This article proposes an extension of the HETOP model, termed the "Fay–Herriot heteroskedastic ordered probit" (FH-HETOP) model due to its ties to small-area estimation (Fay & Herriot, 1979; Ghosh & Rao, 1994; Pfefferman, 2002), as a way to mitigate these shortcomings of the direct estimators. The framework supports joint estimation of $\{\mu_g, \sigma_g\}_{g=1}^{G}$ as well as relationships of these parameters to covariates, using suggestions by Efron (2016) to flexibly model the distributions of μ_g and σ_g across groups. The following section introduces notation for data and the model and discusses options for using the model to estimate group-specific parameters. Results of a simulation study investigating model performance are then presented. This is followed by an example application of the model to real data and a discussion.

The Fay–Herriot HETOP Model

This section first describes the FH-HETOP model. It then discusses distinct applications in which the model may be useful, provides details about specifying and estimating it, and discusses options for using it to estimate group parameters.

Notation and Model Definition

Let $g = 1, \ldots, G$ index groups. Let $Y_{ig} \in \{1, \ldots, K\}$ be the ordinal performance category for individual i in group g. We describe the model as if Y_{ig} is a performance-level category such as "below basic" or "advanced" that would be derived from a score on a standardized assessment, but the modeling framework applies to other circumstances with ordinal data (e.g., Likert-type scale ratings from a survey instrument). It is assumed that Y_{ig} is determined by a latent continuous variable Y^*_{ig} through a vector of $K-1$ "cut points" $\mathbf{c} = (c_1, \ldots, c_{K-1})'$, assumed to be common across groups and satisfying $-\infty < c_1 < \ldots < c_{K-1} < \infty$. Specifically, it is assumed that $Y^*_{ig}|\mu_g, \sigma_g$ are independently and identically distributed (IID) normal random variables with mean μ_g and variance σ^2_g and that $Y_{ig} = 1$ if $Y^*_{ig} \leq c_1$, $Y_{ig} = K$ if $Y^*_{ig} > c_{K-1}$, and $Y_{ig} = k$ if $c_{k-1} < Y^*_{ig} \leq c_k$ for $k = 2, \ldots, K-1$. Thus, Y_{ig} follows the ordered probit model (McCullagh, 1980)

$$\Pr(Y_{ig} = k|\mu_g, \sigma_g, \mathbf{c}) = \begin{cases} \Phi\left(\frac{c_1 - \mu_g}{\sigma_g}\right) \text{ if } k = 1 \\ \Phi\left(\frac{c_k - \mu_g}{\sigma_g}\right) - \Phi\left(\frac{c_{k-1} - \mu_g}{\sigma_g}\right) \text{ if } 2 \leq k \leq K-1 \\ 1 - \Phi\left(\frac{c_{K-1} - \mu_g}{\sigma_g}\right) \text{ if } k = K \end{cases}. \quad (1)$$

This model is assumed to hold for each group, corresponding to a HETOP model because σ_g varies by group. We assume throughout that $K \geq 3$ because the case when $K = 2$ reduces to the standard probit model for a dichotomous outcome, where it generally would not be possible to allow both μ_g and σ_g to vary by group.

The case considered here, as well as by Reardon, Shear, et al. (2017), is when the individual-level data Y_{ig} are not observed. Rather, the observed data are counts $\{N_{gk}\}$ for $g = 1, \ldots, G$ and $k = 1, \ldots, K$, where N_{gk} is the number of individuals from group g who are in ordinal performance category k. Thus, the observed data are both aggregated and coarsened relative to the individual-level, continuous measures Y^*_{ig}. Denote the vector of counts $(N_{g1}, \ldots, N_{gK})'$ for group g by $\mathbf{N}_g$, and let $n_g = \sum_{k=1}^{K} N_{gk}$. Assume that $\mathbf{N}_1, \ldots, \mathbf{N}_G$ are mutually independent conditional on $(\{\mu_g, \sigma_g, n_g\}_{g=1}^{G}, \mathbf{c})$. Then,

$$\mathbf{N}_g|\mu_g, \sigma_g, n_g, \mathbf{c} \overset{\text{Ind.}}{\sim} \text{multinomial}\,(n_g, \boldsymbol{\pi}(\mu_g, \sigma_g, \mathbf{c})), \quad (2)$$

where $\boldsymbol{\pi}(\mu_g, \sigma_g, \mathbf{c})$ denotes the probabilities for the K categories on the right-hand side of Equation 1. The resulting likelihood function may be used to estimate $(\{\mu_g, \sigma_g\}_{g=1}^{G}, \mathbf{c})$ by MLE, provided that some necessary model identification constraints are imposed and that the MLE exists given the observed data. This is the estimation approach advocated by Reardon, Shear, et al. (2017) to obtain the DE $\{\widehat{\mu}_g, \widehat{\sigma}_g\}_{g=1}^{G}$.

Here we extend Model 2 by incorporating covariates and distributional structure of the group parameters. Letting $\mathbf{Z}_g$ be an observed vector of covariates for group g, we define the FH-HETOP model as follows:

$$\begin{aligned}
\mathbf{N}_g|\mu_g, \sigma_g, n_g, \mathbf{c} &\overset{\text{Ind.}}{\sim} \text{multinomial}\ (n_g, \boldsymbol{\pi}(\mu_g, \sigma_g, \mathbf{c})) \\
\mu_g &= \mathbf{Z}_g{}'\boldsymbol{\beta}_\mu + \delta_{\mu,g} \\
\log(\sigma_g) &= \mathbf{Z}_g{}'\boldsymbol{\beta}_\sigma + \delta_{\sigma,g} \\
(\delta_{\mu,g}, \delta_{\sigma,g})'|\boldsymbol{\alpha} &\overset{\text{IID}}{\sim} F(\cdot; \boldsymbol{\alpha}),
\end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_\mu$ is a vector of regression coefficients for the group means, $\boldsymbol{\beta}_\sigma$ is a vector of regression coefficients for the logs of the group SDs, and $F(\cdot; \boldsymbol{\alpha})$ is a bivariate distribution for the residuals $(\delta_{\mu,g}, \delta_{\sigma,g})$ that may depend on additional parameters $\boldsymbol{\alpha}$. Details on the specification of $F(\cdot; \boldsymbol{\alpha})$ are provided in a later section. Certain elements of either $\boldsymbol{\beta}_\mu$ or $\boldsymbol{\beta}_\sigma$ may be set to 0 to allow different subsets of the covariates $\mathbf{Z}_g$ to be included in the mean and log SD models, and $\boldsymbol{\beta}_\mu = \boldsymbol{\beta}_\sigma = \mathbf{0}$ corresponds to the case where no covariates are included in the model. The covariates are general, but in typical applications with achievement data, they may include aggregate student demographic characteristics (such as percentages of students in different racial/ethnic groups and percentages of students participating in free- or reduced price lunch programs) or geographic and economic characteristics of the groups.

The model is analogous to the Fay–Herriot (1979) area-level model used in small-area estimation, where the "areas" in this context are the groups. Modeling the group parameters as a function of both $\mathbf{Z}_g$ and residuals with distribution $F(\cdot; \boldsymbol{\alpha})$ allows the parameters for each group to be informed by both its covariates and information about the heterogeneity of the group parameters among the ensemble of groups. The use of covariates to predict group means is commonplace, given the well-known relationships between aggregate student characteristics and average achievement. The use of covariates to explain SDs is less common though not unprecedented (see, e.g., Gu, Fiebig, Cripps, & Kohn, 2009; Hedeker, Demirtas, & Mermelstein, 2009; Kapur, Li, Blood, & Hedeker, 2015; Kim & Choi, 2008; Leckie, French, Charlton, & Browne, 2014). The assumed linear model for $\log(\sigma_g)$ is analogous to that used by Harvey (1976) to model heteroskedasticity as a function of covariates.

Applications in Which the FH-HETOP Model May Be Useful

There are two distinct applications in which the FH-HETOP model may be useful. The first is for reporting what we will term "derived estimates" of group means and *SD*s suitable for secondary data analysis, which is similar to the goals in some other small-area estimation applications. For example, the Stanford Education Data Archive (Reardon, Ho, et al., 2017) uses the HETOP model to compute estimates of means and SDs of achievement distributions for each school district in the United States, from 2008–2009 through 2014–2015. The raw data used to compute these estimates are aggregate proficiency counts obtained from the Federal Department of Education EDFacts initiative (https://www2.ed.gov/edfacts), which contains these data for every state. There is no equivalent, equally comprehensive database at the national level containing individual-level data. This was the original motivation for Reardon, Shear, et al.'s (2017) application of the HETOP model. A reasonable goal for the derived estimates is that they be suitable for a variety of purposes that are difficult to identify in advance because they depend on the goals of downstream analysts. Applications may include regressions on covariates that may or may not be available during the process of constructing the derived estimates, studies of the distributional properties of group parameters, or studies of nonlinear functions of the parameters such as ICCs. Derived estimates computed from the FH-HETOP model can be particularly effective for estimating some of these parameters as demonstrated in the Simulation Study section.

The other application for which the FH-HETOP model may be useful is when an analyst possesses both group-level covariates and the coarsened, group-level data and is interested in studying the relationships of those covariates to variation across groups in the mean and/or heterogeneity of achievement. Inferences about these relationships can be made from FH-HETOP model by including the covariates of interest in the model so that there is no need to use a two-stage procedure that first obtains DE $\{\widehat{\mu}_g, \widehat{\sigma}_g\}_{g=1}^{G}$ and then relates these estimates to the covariates. We provide an example of such an application in the Empirical Example section.

Specification of $F(\cdot;\boldsymbol{\alpha})$

Implementing the FH-HETOP model requires specifying the residual distribution $F(\cdot;\boldsymbol{\alpha})$. A convenient choice is a Gaussian distribution where $\boldsymbol{\alpha}$ consists of a mean vector and covariance matrix. This assumption is common in applications as well as in software capable of estimating random effects models similar to the FH-HETOP model. However, there are likely going to be applied settings where the joint normality assumption does not hold. Also, joint normality generally will not hold simultaneously for different possible specifications of the covariates $\mathbf{Z}_g$ included in Model 3 due to the changing definition of the

residuals as covariate specifications change. Group parameter estimates obtained from the FH-HETOP model using a bivariate normal specification for $F(\cdot;\boldsymbol{\alpha})$ would be at risk of distortion by the model in cases where normality fails, and misspecification could degrade the properties of estimated regression coefficients as well. Thus, we are interested in less restrictive specifications of the residual distribution.

In the Bayesian literature, the Dirichlet process is commonly used for nonparametric specifications of distributions such as $F(\cdot;\boldsymbol{\alpha})$ (Ferguson, 1973; Ohlssen, Sharples, & Spiegelhalter, 2007; Paddock, Ridgeway, Lin, & Louis, 2006), and Gill and Casella (2009) use such an approach in a setting similar to the FH-HETOP model. Such methods for using nonparametric distributions in Bayesian models are analogous to other nonparametric approaches for specifying latent distributions (Laird, 1978; Lockwood & McCaffrey, 2014; Mislevy, 1984; Rabe-Hesketh, Pickles, & Skrondal, 2003; Roeder, Carroll, & Lindsay, 1996). However, nonparametric specifications can be less efficient than parametric specifications and can introduce problems with model identifiability with discrete data such as those available here (Haberman, 2005). Thus, introducing some degree of smoothness into the specification of $F(\cdot;\boldsymbol{\alpha})$ can be beneficial (Efron, 2014, 2016; Shen & Louis, 1999).

Efron (2016) suggests a class of parametric exponential family distributions that is capable of striking a balance among flexibility, efficiency, and smoothness. To model the distribution of a latent variable δ, such as the residual terms in Model 3, the approach begins by selecting an arbitrarily fine grid $\{\delta_m\}_{m=1}^{M}$ for support of the distribution, and specifying a $(M \times p)$ matrix $\mathbf{Q}$. It then specifies an exponential family probability distribution on the grid with probabilities $f(\delta_m;\boldsymbol{\alpha}) = \exp\{\mathbf{Q}_m\boldsymbol{\alpha} - \varphi(\boldsymbol{\alpha})\}$, where $\boldsymbol{\alpha}$ is a p-dimensional vector of unknown parameters, $\mathbf{Q}_m$ is row m of $\mathbf{Q}$, and $\varphi(\boldsymbol{\alpha})$ is a normalizing constant to make the probabilities sum to 1. Thus, rather than letting the probabilities on the grid be unconstrained, which would require $M-1$ free parameters, the probabilities are constrained to follow the specified functional form that depends on only p parameters. In practice, $p << M$ would be selected for parsimony, but flexibility in the shape of the distribution can be maintained through the specification of $\mathbf{Q}$. For example, letting the p columns of $\mathbf{Q}$ be a cubic B-spline basis (e.g., Eilers & Marx, 1996) allows the curve connecting the logs of the probabilities to be a continuous, piecewise cubic polynomial which is capable of capturing a wide variety of distributional shapes. We refer to the specification of M, the grid, and the matrix $\mathbf{Q}$ as an "Efron prior" with p-dimensional parameter $\boldsymbol{\alpha}$. The Efron prior is analogous to the "Ramsay curve" method of specifying latent ability distributions in IRT models (Monroe & Cai, 2014; Woods & Thissen, 2006). The introduction of a discrete grid is for computational convenience rather than mathematical necessity, and the supporting theory carries over to the continuous case (Efron, 2016). By approximating the latent distribution with a discrete

distribution, the approach is similar to some nonparametric distribution specifications; however, the functional form constraint on the probabilities can be used to force those probabilities to vary more smoothly across the grid. Efron (2016) provides empirical examples that demonstrate the value of this model for being sufficiently flexible to capture complicated structure in the latent distribution, while simultaneously being sufficiently constrained to mitigate estimation error. We view these benefits as being ideally suited to the FH-HETOP case, and to our knowledge this application is novel to small-area estimation settings.

A challenge to implementing the Efron prior in the context of the FH-HETOP model is that the model requires a bivariate specification $F(\cdot;\boldsymbol{\alpha})$, whereas Efron (2016) directly considers only the univariate case. Extension to the bivariate case would be possible using bivariate splines but would entail computational difficulties with fine grids. Thus, we opt for a simpler specification with a univariate Efron prior for one dimension, and then a second univariate Efron prior for the residual of the second dimension given the first. We use a linear regression to allow for correlation between the dimensions, which is likely to be a suitable approximation in many circumstances, and can be expanded if more flexibility is needed. Details are provided in the Appendix in the online version of the journal.

Identification and Model Estimation

The locations and scales of the parameters $\{\mu_g\}_{g=1}^G$ and $\{\sigma_g\}_{g=1}^G$ are indeterminate without additional identification assumptions (see Reardon, Shear, et al., 2017, for an extensive discussion). We opt to identify the locations and scales of these parameters by fixing exactly two of the cut points. This is sufficient for identification for any $K \geq 3$.[2] In cases with covariates $\mathbf{Z}_g$, we exclude an intercept and center each covariate to have mean 0 across groups. This allows any nonzero means of μ_g or $\log(\sigma_g)$ to be absorbed by the residual distribution $F(\cdot;\boldsymbol{\alpha})$, which is simpler to implement than alternative specifications in which the mean of $F(\cdot;\boldsymbol{\alpha})$ is constrained to be 0.

Given a complete model specification, including identification constraints and a specification for $F(\cdot;\boldsymbol{\alpha})$, there are two common estimation approaches for models such as the FH-HETOP model: an empirical Bayesian approach (Carlin & Louis, 2000; Casella, 1985; Efron & Morris, 1973; Morris, 1983), and a fully Bayesian approach (see, e.g., Gelman, Carlin, Stern, & Rubin, 1995). In the empirical Bayesian approach, the model parameters $(\mathbf{c}, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\alpha})$ would be estimated by maximum likelihood or related methods, while the group parameters $\{\mu_g, \sigma_g\}_{g=1}^G$ would then be estimated conditional on $(\widehat{\mathbf{c}}, \widehat{\boldsymbol{\beta}}_\mu, \widehat{\boldsymbol{\beta}}_\sigma, \widehat{\boldsymbol{\alpha}})$, typically using posterior means (PMs) or modes. In the fully Bayesian approach, $(\mathbf{c}, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\alpha})$ would be given prior distributions, and a posterior distribution for these parameters and the group parameters $\{\mu_g, \sigma_g\}_{g=1}^G$, given the observed data would be obtained, often via Markov Chain Monte Carlo (MCMC) methods

(Gilks, Richardson, & Spiegelhalter, 1996). The application of MCMC to ordinal data settings similar to the FH-HETOP model is discussed by Albert and Chib (1993); DeYoreo and Kottas (2017); Johnson (1996); Johnson and Albert (1999); Lockwood, Savitsky, and McCaffrey (2015); Savitsky and McCaffrey (2014); and Segawa (2005).

We opt for the fully Bayesian approach for several reasons. First, it has the advantage of automatically incorporating uncertainty about $(\mathbf{c}, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\alpha})$ into inferences about group parameters because the marginal posterior distribution of the group parameters integrates over $(\mathbf{c}, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\alpha})$. The fully Bayesian approach using MCMC is also attractive because the posterior samples provide an automatic way of providing inferences for arbitrarily complicated functions of the group parameters such as the ICC, distribution functions, and percentiles (Paddock & Louis, 2011). Finally, the fully Bayesian approach supports straightforward implementation of several options for constructing estimates of the group parameters $\{\mu_g, \sigma_g\}_{g=1}^G$ from MCMC samples, as discussed in the following section. Details on our fully Bayesian specification of the FH-HETOP model with Efron priors are provided in the Appendix in the online version of the journal.

Group Parameter Estimation

Special considerations are needed when the FH-HETOP model is used to generate derived estimates of group parameters because under the model, $\{\mu_g, \sigma_g\}_{g=1}^G$ depend in part on random effects distributed according to $F(\cdot; \boldsymbol{\alpha})$. PMs are common estimators in either the empirical or fully Bayesian model. Such estimators are optimally accurate in terms of mean squared error (MSE), but in contrast to the overdispersion of the DE, PMs are underdispersed relative to the true distributions of the group parameters (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Shen & Louis, 1999; Warm, 1989). This makes PMs unattractive for some uses of derived estimates because their empirical distribution across groups does not well approximate the corresponding distribution of the true group parameters, and PMs do not have proper covariances with any covariates that were not used in their construction.

Here, we consider two options that have been developed to solve the problem of underdispersion of PMs. The first is so-called constrained Bayes (CB) estimators (Devine & Louis, 1994; Ghosh, 1992; Louis, 1984), which rescale PMs of $\{\mu_g\}_{g=1}^G$ and $\{\sigma_g\}_{g=1}^G$ so that each ensemble has variance equal to the estimated marginal variance of the latent parameters. This tends to reduce the amount of shrinkage bias in the PMs. The second option we consider is "triple goal" (TG) estimators (Paddock et al., 2006; Shen & Louis, 1998, 2000), which aim to be simultaneously effective for point estimation, estimation of ranks, and estimation of the distribution of the latent parameters. For example, the TG estimates of

$\{\mu_g\}_{g=1}^G$ are computed by first estimating the PM of the rank of each μ_g across the groups. These are then stretched to be equally spaced percentile ranks $\widehat{r}_g$ on (0, 1), and then the TG estimates $\{\widehat{\mu}_g\}_{g=1}^G$ are defined as the inverse of the estimated true CDF of $\{\mu_g\}_{g=1}^G$ evaluated at $\widehat{r}_g$. The TG estimators provide MSE-optimal rank estimates, their histogram reasonably approximates the true latent distribution (i.e., they are neither underdispersed nor overdispersed), and they tend to demonstrate both small bias and small MSE for the individual group parameters as well (Shen & Louis, 1998). Thus, either the CB or TG estimators for $\{\mu_g\}_{g=1}^G$ and $\{\sigma_g\}_{g=1}^G$ may be a reasonable way to compute derived estimates using the FH-HETOP model that may function well for a variety of secondary data analyses.

Simulation Study

We conducted a simulation study focused on the performance of derived estimates of $\{\mu_g\}_{g=1}^G$ and $\{\sigma_g\}_{g=1}^G$. The goal was to compare the DE obtained by MLE as well as PMs, CB estimators and TG estimators from the FH-HETOP model, in terms of their performance for several inferential goals representative of how derived estimates may be used in applications. All simulations used $K = 4$. A total of 32 simulation conditions were examined, obtained by crossing two choices for the number of groups ($G = 200$ or 400), four choices for the group sample sizes ($n_g = 12, 25, 50$, or 100), and four choices for the cut point locations $\mathbf{c}$ corresponding to marginal category probabilities of $(0.25, 0.25, 0.25, 0.25)$, $(0.05, 0.45, 0.45, 0.05)$, $(0.05, 0.25, 0.25, 0.45)$, or $(0.05, 0.10, 0.65, 0.20)$. We chose to vary these three factors (G, n_g and $\mathbf{c}$) because each can affect the amount of information that the observed data provide about the unknown parameters. For example, data, where both G and n_g are large, and the categories are about equally frequent, are more informative than data, where G and n_g are relatively small, and one or more categories is rare. The focus on relatively modest sample sizes per group was motivated by the facts that (a) the candidate estimators may behave differently from one another in such cases, whereas they will be similar when groups have large samples; and (b) small samples are likely to be encountered in many practical applications (e.g., the median n_g in the data used in the Empirical Example section is 12). For each of the 32 simulation conditions, 100 independent replications were conducted for a total of 3,200 replications.

A replication of the simulation for a given value of G, n_g, and $\mathbf{c}$ proceeded as follows. First, scalar covariates $\{Z_g\}_{g=1}^G$ were generated as $(1/\sqrt{2})$ times IID draws from a Student's t distribution with 4 degrees of freedom so that the covariates have mean 0 and variance 1. Then, $\{\mu_g\}_{g=1}^G$ were generated as a linear

regression on $\{Z_g\}_{g=1}^G$ with residuals IID from a scaled and centered χ_1^2 distribution, and $\{\log(\sigma_g)\}_{g=1}^G$ were generated as a linear regression on $\{Z_g\}_{g=1}^G$ with residuals IID from a scaled and centered χ_1^2 distribution, and independent of the residuals for $\{\mu_g\}_{g=1}^G$. The regression coefficients were selected so that the covariate had $R^2 = .50$ for the means and $R^2 = .10$ for the log *SD*s. The generated true parameters $\{\mu_g, \sigma_g\}_{g=1}^G$ were then transformed to an alternate scale $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$ consistent with a population mean of 0 and population *SD* of 1; details are in the Appendix in the online version of the journal. Across groups and simulation replications, μ_g^* has mean = 0 and $SD = .48$, while σ_g^* has mean = .86 and $SD = .17$. To complete the data generation, cut points $\mathbf{c}$ were selected to achieve one of the four aforementioned scenarios for the marginal category probabilities, and then count data $\{N_{gk}\}$ were generated from the appropriate group-specific multinomial distributions with $n_g \in \{12, 25, 50, 100\}$, depending on the simulation condition.

For each simulation replication, the simulated count data $\{N_{gk}\}$ were used to compute the DE as the MLE of $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$.[3] Arbitrary identification rules are required to ensure the existence of the MLE. We adopted the following rules for a given set of simulated count data $\{N_{gk}\}$. For any group g with nonzero counts in fewer than three of the four categories, $\log(\sigma_g)$ was constrained to equal the mean of $\log(\sigma_g)$ for the remaining groups. For example, if in a given simulated data set with $G = 200$, six groups had nonzero counts in fewer than three categories, $\log(\sigma_g)$ for each of these six groups was constrained to equal the mean of $\log(\sigma_g)$ for the other 194 groups. In addition, constraints were imposed on the mean parameters for groups in which all data fell into either the lowest or highest category. Let $\mathcal{G}$ be the set of groups for which it is not the case that all data fall into an extreme category. Then, for any group g with all data in the lowest category, we set $\mu_g = \min_{g' \in \mathcal{G}}(\mu_{g'})$, and similarly for any group g with all data in the highest category, we set $\mu_g = \max_{g' \in \mathcal{G}}(\mu_{g'})$. Collectively, these constraints reduce the dimension of the parameter space to force existence of the MLE by assigning unidentified parameters to values informed by groups with better data.

After obtaining the DE, we used $\{N_{gk}\}$ to estimate the FH-HETOP model in Just Another Gibbs Sampler (JAGS) (Plummer, 2003). The JAGS model code for the FH-HETOP model, as well as a brief description of an accompanying package "HETOP" for the R environment that we developed to implement all estimators used in this article, is given in the Appendix in the online version of the journal. The first two cut points were fixed at -1 and 0, respectively, and the Efron priors used $\mathbf{Q}$ with $M = 100$ grid points equally spaced from $[-5, 5]$ for the means and $[\log(0.10), \log(5.0)]$ for the log *SD*s, and cubic B-splines with $p = 10$ degrees of freedom. Five thousand posterior samples were collected from each of two independent chains after 2,000 burn-in iterations, and convergence

was verified with Gelman-Rubin (1992) statistics. The samples of group parameters were transformed, iteration by iteration, using the transformation function in the Appendix in the online version of the journal that puts the estimates on the scale consistent with $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$. Importantly, the covariate Z_g for each group was omitted from the estimation of the FH-HETOP model. This puts the DE and the estimates from the FH-HETOP model on equal footing, allowing us to compare their performance in a secondary regression model with a covariate that was not used in either model. A summary of additional simulation results in which Z_g was included in the FH-HETOP model is provided near the end of this section.

The estimation process for a simulation replication resulted in four sets of estimates $\{\widehat{\mu}_g^*, \widehat{\sigma}_g^*\}_{g=1}^G$: the DE and then PMs, CB, and TG from the FH-HETOP model. These four sets of estimates were compared with respect to their performance for estimating $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$ and their distributions, the between-group ICC of achievement, and the regression of the group parameters on the omitted covariate Z_g. Each of these sets of results is discussed in turn. Results are summarized by pooling across all 3,200 replications for the 32 simulation conditions, with any key findings for particular conditions noted.

Results: Group Parameter Estimation

The top half of Figure 1 summarizes the bias (left) and root mean squared error (RMSE; right) for the different estimators of the group means. The true group means μ_g^* were sorted and broken into 20 bins each containing 5% of the distribution, and the figures provide the bias and RMSE of each estimator by bin. The population *SD* of μ_g^* is .48, which can be used to calibrate the magnitude of the bias and RMSE of the mean estimators with respect to the population distribution of the true means. The bias for DE is smallest overall, while the other estimators demonstrate the expected shrinkage bias, which is most severe for PM. However, the empirical distribution of DE across groups and simulation replications has *SD* = .53, which is overdispersed relative to the true parameters and contributes to the lower accuracy of DE relative to the other estimators (top right of Figure 1). The empirical distribution of PM is underdispersed, with *SD* of .45. Alternatively, both the TG and CB estimators have *SD* of .48, matching the truth. The first row of Table 1 shows that the overall accuracy of CB and TG is close to that of PM, which is optimal in terms of RMSE.

The bottom half of Figure 1 is analogous to the top half, but is for the group SDs. DE has negative bias across the range of true parameters, whereas the other estimators again demonstrate the expected shrinkage bias, with CB and TG demonstrating somewhat less such bias than PM. The empirical distribution of DE is substantially overdispersed relative to the true distribution of σ_g^*, leading to lower accuracy across the distribution of σ_g^* (bottom right of Figure 1) and thus

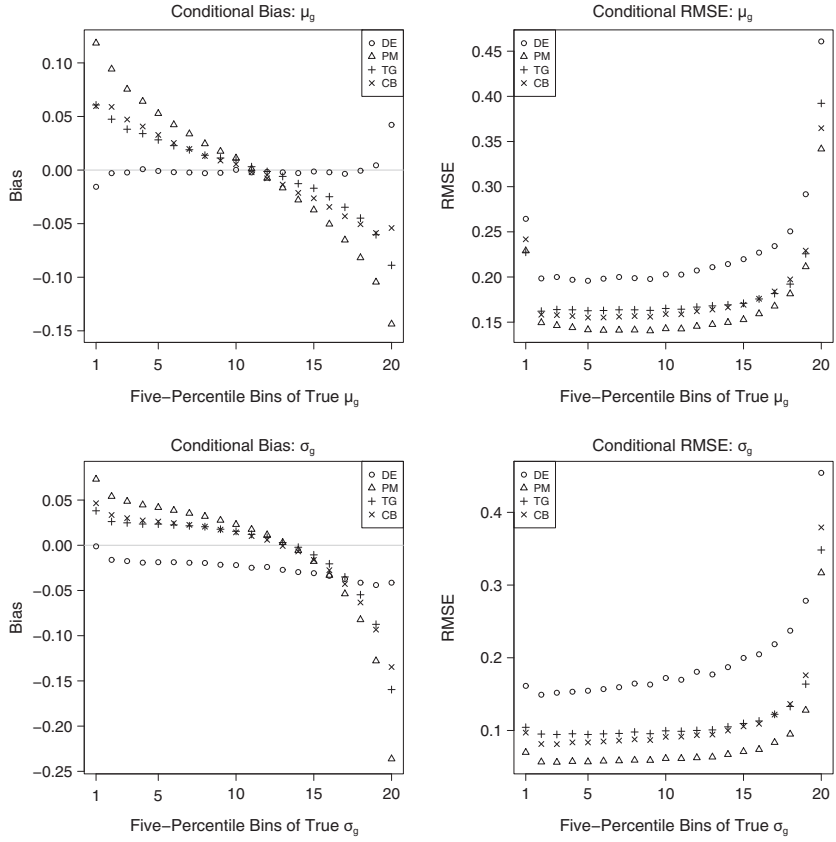

FIGURE 1. *Bias and root mean squared error for different estimators of group means (top half) and group standard deviations (bottom half) from the simulation study, conditional by 5 percentile bins of the corresponding true parameters.*

overall (second row of Table 1). Again the CB and TG estimators are competitive with PM in terms of accuracy. It is worth noting that all estimators demonstrate overall lower accuracy for estimating σ_g^* than they do for estimating μ_g^*, as the magnitudes of the RMSEs relative to the variation of the true parameters are notably larger for σ_g^* (population $SD = .17$) than they are for μ_g^* (population $SD = .48$). This suggests the coarsened data provide relatively weaker information about within-group variability of achievement than they do about within-group level of achievement.

The basic patterns of the performance of the different estimators for μ_g^* and σ_g^* were largely insensitive to the simulation condition though accuracy naturally

TABLE 1.
Root Mean Squared Errors of Derived Estimators for Different Target Estimands From Simulation Study

Target	Direct Estimates	Posterior Means	Constrained Bayes	Triple Goal
$\{\mu_g^*\}_{g=1}^G$	.236	.185	.191	.196
$\{\sigma_g^*\}_{g=1}^G$	.208	.122	.139	.139
Intraclass correlation coefficient	.054	.033	.027	.027
Regression of means on $\{Z_g\}_{g=1}^G$	.025	.064	.040	.041
Regression of log(standard deviations) on $\{Z_g\}_{g=1}^G$	.021	.030	.030	.026

was larger when n_g was larger and/or the category frequencies more balanced. There was negligible sensitivity to G.

Results: ICC

The between-group ICC of the latent variable is defined as the ratio of the between-group variance to the population variance. In the parameterization $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$ that imposes population variance of one, the ICC is simply the variance of $\{\mu_g^*\}_{g=1}^G$. Not surprisingly, the overdispersion of the DE leads to notable positive bias in the estimated ICC. The solid curve in Figure 2 is the estimated density, across the 3,200 simulation replications, of the difference between the ICC computed from the DE and the true ICC based on $\{\mu_g^*\}_{g=1}^G$. The corresponding densities for the PM and TG estimates are provided with other line types, with the density for CB omitted because it is nearly identical to that of TG. The ICC estimated from the DE is positively biased with a heavy right tail. The ICC estimated from PM has a smaller negative bias, whereas the ICC from the TG (and CB) estimates has almost no bias and is comparatively precise. Row 3 of Table 1 summarizes the higher accuracy of the ICC estimated from either TG or CB relative to the alternatives.

Results: Regression on Z_g

Recall that the true parameters were generated with a regression relationship on Z_g, but that Z_g was not used in the construction of the derived estimates. This allows us to examine the performance of the derived estimates in a second-stage regression model, as might be common in applications using derived estimates for secondary analysis involving covariates not used or available during the computation of the derived estimates. For the group means, we define the "true" regression coefficient by the linear regression of $\{\mu_g^*\}_{g=1}^G$ on $\{Z_g\}_{g=1}^G$. This varies

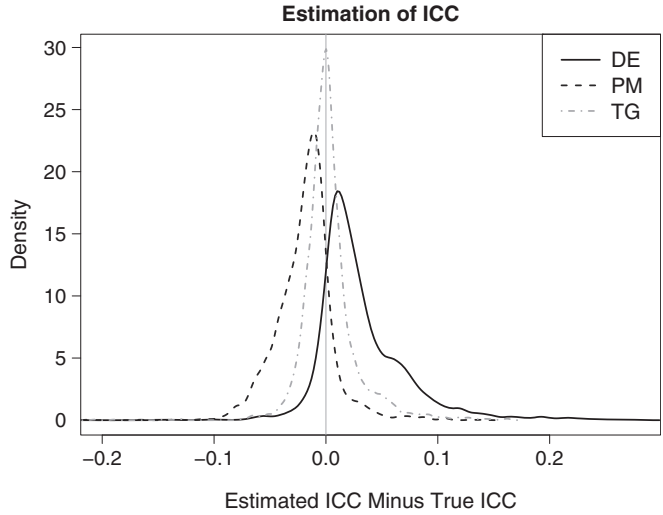

FIGURE 2. *Estimated densities, across simulation replications, of the difference between the estimated and true intraclass correlation coefficients. Constrained Bayes is almost identical to triple goal and is omitted from the figure. Vertical line at 0.*

across simulation replications due to the random generation of both the covariates and the group parameters. Across all 3,200 simulation replications, the true regression coefficient had mean = .34 and *SD* = .02. The solid curve in the top left frame of Figure 3 is the estimated density, across all simulation replications, of the difference between the coefficient obtained from regressing the DE of the group means on Z_g and the true regression coefficient. The corresponding densities for the PM and TG estimates are provided with other line types, with the density for CB omitted because it is nearly identical to that of TG. Second-stage regression using DE is approximately unbiased and is most accurate (fourth row of Table 1), while that using PM has a large negative bias due to shrinkage, which degrades accuracy. These problems are partially mitigated by second-stage regression with TG.

The top right frame of Figure 3 is analogous to the top-left frame but is instead for regressions of the log group *SD*s on Z_g. Analogous to the means, the "true" regression coefficient for a simulation iteration is defined by the linear regression of $\{\log(\sigma_g^*)\}_{g=1}^G$ on $\{Z_g\}_{g=1}^G$. This coefficient has mean = .05 and *SD* = .01 across simulation iterations. The patterns are similar to those for the group means though in this case the density for CB (not shown) is shifted slightly below that of TG, leading to somewhat lower accuracy of CB relative to TG (fifth row of Table 1).

Variations in performance by specific simulation condition were predictable: All estimators were more accurate with larger n_g, more balance in category

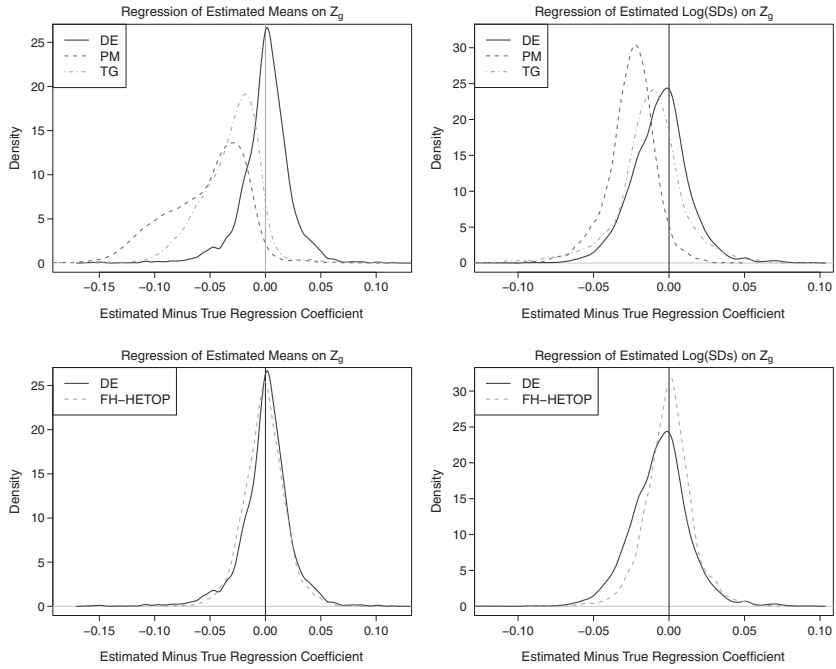

FIGURE 3. *Estimated densities, across simulation replications, of the difference between the estimated and true regression coefficient of the group means (top left frame) and group log standard deviations (top right frame) on* Z_g. *The bottom frames are analogous to the top frames, but instead compare direct estimates (DE; reproduced from the top frames) to the posterior means of the regression coefficients obtained by including* Z_g *in the Fay–Herriot heteroskedastic ordered probit (FH-HETOP) model. Vertical line at 0.*

frequencies, and larger G, all of which correspond to having more information to infer regression relationships. This held for both regressions of estimated means on the covariate, and regression of the estimated log SDs on the covariate. In both cases, the loss of accuracy due to shrinkage was most pronounced for $n_g = 12$, but even in that case, TG was notably more accurate than PM.

The bias for the regression coefficients estimated using derived estimates from the FH-HETOP model can be eliminated by including the covariate in the model. To demonstrate, we conducted a parallel set of simulations that fit the FH-HETOP model with Z_g included in the model and computed the PM of the regression coefficient from both the mean and log SD regression models after transforming those coefficients to the scale appropriate for $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$ (see Appendix in the online version of the journal). The bottom frames of Figure 3 provide the same densities for DE as the top frames, but now compare these densities to the corresponding densities for the regression coefficients estimated

from the FH-HETOP model with Z_g included in the model. Estimating the regression coefficients directly from the FH-HETOP model eliminates the bias, and it also improves accuracy compared to estimating the regression coefficients using a two-stage procedure. The RMSEs for the true regression coefficients from the FH-HETOP model are .019 and .016 for the mean and log SD models, respectively, which improve upon all of the estimators in Rows 4 and 5 of Table 1, and specifically are about 24% smaller than the RMSEs achieved with the DE.

Summary of Simulation Results

The simulation study demonstrates that the FH-HETOP model can produce derived estimates that can be useful for various inferential goals. The CB and TG estimates largely live up to their promise of providing a single set of estimates that are suitable for many purposes and tend not to perform poorly even in cases where they are not optimal (e.g., for estimating regression coefficients for covariates not used during the estimation). These benefits result primarily from the fact that their empirical distribution across groups is designed to match features of the corresponding distribution of the true parameters. This comes at the price of some degree of shrinkage bias in their covariances with omitted variables, but this bias is not as severe as it is for PMs. The simulation also demonstrates that this bias can be eliminated, and accuracy improved, by using the FH-HETOP to directly model relationships between group parameters and covariates without using a two-stage procedure.

We also conducted a set of simulations parallel to those presented here, but using $p = 5$ degrees of freedom for the Efron priors rather than $p = 10$. The results were extremely similar. The only difference worth noting is that the bias in the estimates of the group *SD*s using $p = 5$ was slightly smaller for small values of the true group *SD*s, which led to slightly improved performance for the regression coefficients for the group log *SD*s. Model selection criteria can be used to select p. An example is provided in the following section.

Empirical Example

As previously noted, the FH-HETOP can support either the production of derived estimates or the direct estimation of relationships of covariates to group parameters. The simulation study focused primarily on the former. This section presents an example of the latter, using a subset of the Early Childhood Longitudinal Study—Kindergarten cohort (ECLS-K) data set to demonstrate the use of the FH-HETOP model to study the relationships between school-level covariates and school-level means and SDs of math proficiency. These data are publicly available and thus can be further explored with the accompanying "HETOP" R package (described in Appendix in the online version of the journal).

Data Description and Context

We merged the "ecls_child.dta" and "ecls_school.dta" data sets available through the website (http://www.stata-press.com/data/mlmus3.html) for the textbook *Multilevel Modeling Using Stata, Volume II* (Rabe-Hesketh & Skrondal, 2012), and described on pages 626–627. The outcome variable of interest is the math proficiency ("profmath") variable, which is coded as an ordinal variable taking on values 0 to 5. The ordinal categories represent the "highest developmental milestone [in math] reached by the child" (Rabe-Hesketh & Skrondal, 2012, p. 626) at the end of Kindergarten, as determined by the student's performance on items within item clusters that correspond to each developmental milestone (e.g., Milestone 1 is "number and shape" and Milestone 2 is "relative size"). These proficiency levels are thus distinct from those typically reported in K–12 standardized testing that is defined by cut scores along the score scale determined by a standard-setting panel. However, for illustration of the FH-HETOP model, we can still posit that there is a continuous, latent unidimensional math ability construct that is measured by a coarsened ordinal outcome using the items and their scoring procedure.

For this illustration, we assume we are interested in relationships between covariates and the school-level means and *SD*s. We sum the number of students at each proficiency level to compute the counts N_{gk}. Given that only 9 of the 6,477 students obtained "profmath" of 0 (i.e., did not pass any developmental milestone), we combine categories "0" and "1," resulting in $K = 5$. There are $G = 569$ schools with n_g ranging from 2 to 22 (median $= 12$). The preponderance of small schools makes this a prime data set for a FH-HETOP model application, given Reardon, Shear, et al.'s (2017) findings of negative bias for DE $\widehat{\sigma}_g$ for small groups. Moreover, 117 of the 569 schools (20.6%) have nonzero counts for only one or two of the $K = 5$ categories, so that the DE computed by MLE do not exist for these data.

The data set includes several student- and school-level covariates that can be used to model covariate relationships with the school means and *SD*s. We focus on three student-level covariates, which we aggregate to the school level, and two school-level covariates. These covariates are:

- Mean number of student risks (mnNumRisks): Average number of student risks recorded for each student out of four possible risks;
- Mean student socioeconomic status (MnSES): Average student SES (continuous) composite of five standardized measures (details available from National Center forEducation Statistics, 2002, pp. 7–25);
- Percentage of male students (percMale): Percentage of male students in the school;
- Neighborhood climate index (Nbhoodcl): School-level covariate that reflects the principal's perception of six specific problems in the school neighborhood (takes on integers from 0 to 12, with larger values indicating more perceived problems); and

- Private school indicator (Private): School-level covariate that is coded as 1 for private schools and 0 for public schools.

An analyst may want to model the covariate relationships with the group means and *SD*s to identify what type of schools have high/low mean math proficiency and high/low variance. High mean/low variance schools could be considered ideal. However, schools with high mean math proficiency and high variance would also be useful to identify, as the high mean indicates that overall they are performing well, while the high variance indicates that the school may not be serving all students.

Method

To determine the value of using covariates to explain variation in μ_g and σ_g and to illustrate the process for selecting p for the Efron prior, we fit models with and without the covariates and varied the value of p used to define the matrix $\mathbf{Q}$ across $p = 3, 4, 5, \ldots, 14, 15$. The number of rows M of $\mathbf{Q}$ should be chosen to be as large as computationally tolerable, and we fixed $M = 100$. We identified the model by fixing the first two cut points at -2 and -1. We specified the range of the grid for the residual terms iteratively, trying certain grids and then evaluating the distribution of residuals to determine if the upper or lower boundary points needed to be readjusted. We found that grid ranges of $[-2, 7]$ for the mean residuals and $[\log(0.10), \log(5.0)]$ for the log *SD* residuals were sufficient to support the distribution. For each p and model type (with and without covariates), we ran two independent chains with 2,000 burn-in iterations and 5,000 iterations saved for inferences. Convergence was verified with Gelman and Rubin (1992) statistics. For each p and model type, we computed the Watanabe–Akaike information criterion (WAIC; Vehtari, Gelman, & Gabry, 2017), a model selection criterion. Vehtari, Gelman, and Gabry (2017) argue that WAIC improves upon the deviance information criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002), commonly used for selection among Bayesian models, in terms of computational stability and parameterization invariance. Smaller values are preferred.

Results

To determine the effect of including the covariates on model fit and the ideal choice of p, we compare the WAICs, shown in Figure 4. The figure shows that the models that include the covariates (dashed line) fit notably better than those that do not, and that the WAICs fluctuate for even and odd values of p due to the location of the knots for the spline basis, but are fairly stable for larger values from $p = 10$ to $p = 15$. These results support using the model with covariates, which has smallest WAICs for $p = 4, 6,$ and 8. Closer inspection of the WAIC values indicate that the WAICs for $p = 6$ and $p = 8$ are very similar and slightly

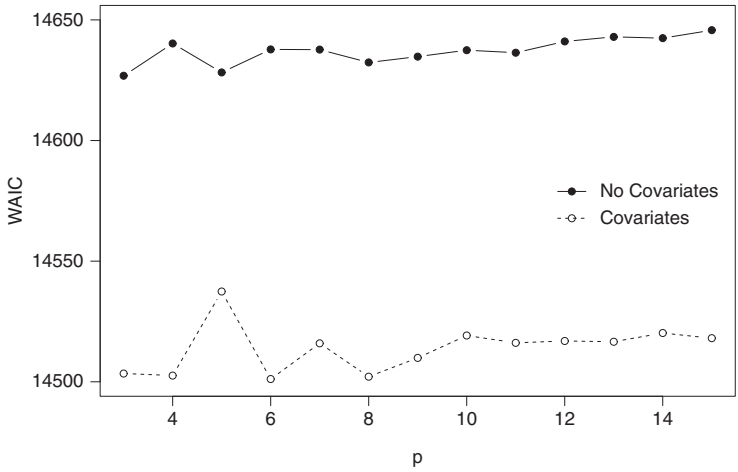

FIGURE 4. *Watanabe–Akaike information criterions to compare model fit in the empirical data analysis.*

smaller than for $p = 4$. The same conclusions were supported by an independent replication of the entire analysis, conducted to ensure that results were not sensitive to Monte Carlo error in the WAIC estimates.

We are also interested in how sensitive inferences about the regression coefficients are to the model specifications. Figure 5 plots the 2.5th and 97.5th percentiles of the posterior samples of the estimated regression coefficients, with the PMs indicated by an asterisk, for each covariate in the modeling of the group means and log *SD*s. PMs and intervals are provided for each value of p. The regression coefficients are presented on the scale where the latent math proficiency has population mean $= 0$ and variance $= 1$. Thus, coefficients from the model for the school means can be interpreted as population *SD* units of proficiency associated with one-unit changes in the school-level covariates. Alternatively, coefficients from the model for the log school *SD*s that are not far from 0 in absolute value can be interpreted as percent changes in the within-school *SD* of proficiency associated with one-unit changes in the school-level covariates. The PMs and intervals are generally quite stable across all values of p, with the most sensitivity tending to occur for the very small values of p that are ruled out by the WAIC.[4] Table 2 provides the PMs and *SD*s of the regression coefficients for each covariate for the model with $p = 6$, the more parsimonious of the two models ($p = 6, 8$) preferred by the WAIC. The table also provides the 2.5th and 97.5th percentiles of the posterior samples. The asterisked covariates are those where the 95% credible interval does not contain 0. The mean number of student risks, mean student SES, and neighborhood climate all are significantly related to school mean performance. Only one covariate had a significant relationship with

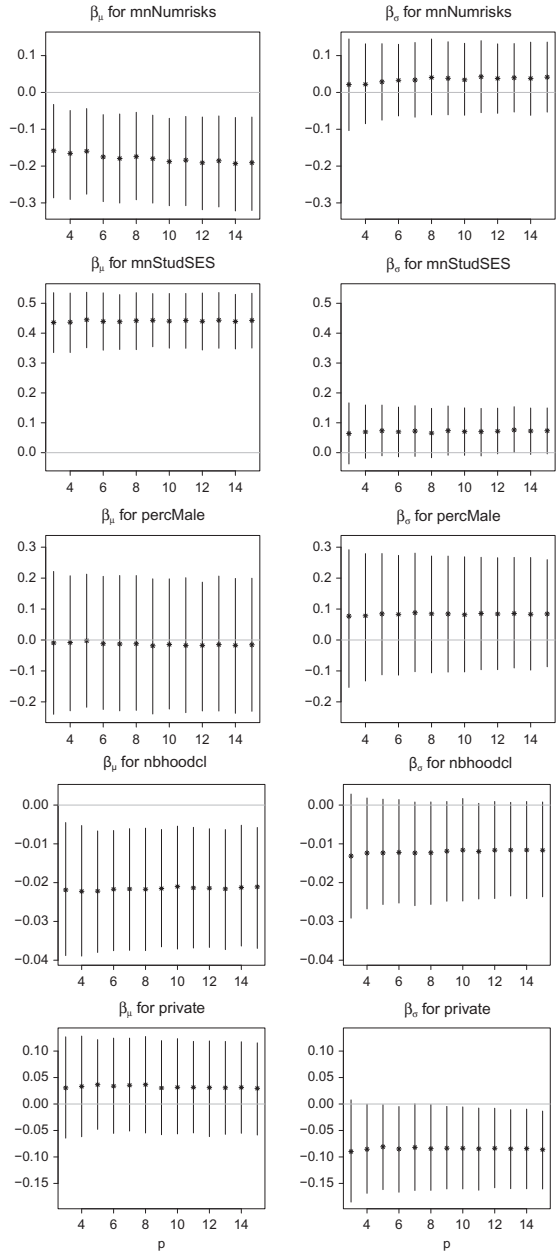

FIGURE 5. *Ninety-five percent credible intervals and posterior means (stars) for estimated regression coefficients by* p.

TABLE 2.
Summary Statistics of Estimated Regression Coefficients From the Empirical Example ($p = 6$)

Parameter	Covariates	Posterior Mean	Posterior Standard Deviation	2.5th Percentile	97.5th Percentile
μ_g^*	mnNumrisks*	−.18	.06	−.30	−.06
	mnStudSES*	.44	.05	.34	.53
	percMale	−.01	.11	−.22	.21
	nbhoodcl*	−.02	.01	−.04	−.01
	Private	.03	.04	−.06	.12
σ_g^*	mnNumrisks	.03	.05	−.06	.13
	mnStudSES	.07	.04	−.01	.15
	percMale	.08	.10	−.11	.27
	nbhoodcl	−.01	.01	−.03	.00
	Private*	−.08	.04	−.17	.00

Note. The 97.5th percentile for nbhoodcl for σ_g^* is +.0014, while that for private is −.0048. mnNumRisks = mean number of student risks; MnSES = mean student socioeconomic status; percMale = percentage of male students; Nbhoodcl = neighborhood climate index; private = private school indicator.

school variance: Private schools, on average, have less variance in their students' math proficiency than public schools. This may result from private schools tending to serve more homogeneous populations than public schools, but further investigation of the schools would be required to explore that hypothesis.

We also used the model output to compute the posterior distribution of the adjusted R^2 of the covariates for the means and log *SD*s. For the means, the PM adjusted R^2 is .67 with 95% credible interval of $(0.57, 0.77)$, whereas for the log *SD*s, the PM adjusted R^2 is .04 with 95% credible interval of $(0.01, 0.09)$. Not surprisingly, the covariates are much more effective at explaining variation in the group means but appear to have some predictive value for the group *SD*s as well.

Discussion

Even though individual-level achievement data are increasingly archived, obtaining such data can be difficult due to increasing privacy concerns, and processing such data can be costly when inferences are needed for many jurisdictions or reporting agencies. The FH-HETOP model may be valuable in such cases because it relies only on aggregate data that are easier to obtain and process. Also, the model can be used in other ordinal data settings where continuous data generally would be unavailable, such as with Advanced Placement® scores or Likert responses to survey instruments. The model also can be extended straightforwardly into a multilevel modeling framework if individual-level

ordinal data and associated covariates, rather than group-level aggregates, are available (see Gu et al., 2009, for such an extension in a setting similar to the FH-HETOP model). In such cases, there may be covariates at both the individual and group levels, providing the opportunity to study contextual effects separately from individual-level effects. The Efron priors for random effects may be useful to provide a multilevel estimation framework that provides more flexibility than the standard assumption of multivariate normality for the random effects. Future work could consider the properties of such estimators.

In settings where the FH-HETOP model is used to report derived estimates of group parameters, we focused on the case where only a single set of derived estimates is computed. In other similar applications, such as with skill estimates obtained from large-scale achievement surveys, multiple sets of derived estimates or "plausible values" are reported (Mislevy et al., 1992). Like the CB and TG estimates, these can mitigate the shrinkage bias of PMs, and the fully Bayesian version of the FH-HETOP model could be easily used to generate appropriate sets of plausible values for secondary analysis.

However, the simulations demonstrate that it is impossible for any one method for producing derived estimates to be best for all possible secondary analysis purposes. The DE has notable deficiencies for some inferences (e.g., distributional properties and ICC), whereas the TG and CB estimates perform reasonably well for most inferences but have more bias and error than the DE for regressions on covariates not used in their construction. It thus may be sensible to report multiple types of derived estimates, with guidance about the most suitable uses for each type, so that secondary analysts can use the estimates that are best aligned with their inferential goals.

In general, there are pros and cons to using covariates in the construction of derived estimates. Using covariates may improve the suitability of derived estimates for some secondary data analysis purposes, such as regression modeling. However, there may be fairness arguments against using covariates in the derived estimates because it implies that two groups with the same observed counts will get different estimates depending on group attributes. The FH-HETOP model can provide useful derived estimates whether or not covariates are used but whether to use covariates at all may require careful consideration in some settings. If covariates will be used to compute derived CB estimates, Kubokawa and Strawderman (2013) and Lyles, Moore, Manatunga, and Easley (2009) discuss methods for constraining the estimates to achieve target covariances with covariates.

While we focused on the FH-HETOP model, we also suggest, and evaluate via simulation, a procedure for direct estimation that ensures the existence of a finite MLE outside of extreme boundary cases (such as when all groups have sparse data). Such a procedure was not provided by Reardon, Shear, et al. (2017) although they consider parameter constraints designed to address small-sample biases. Our procedure that uses constraints for groups with unidentified parameters due to sparse count data builds on that work and provides analysts

interested in direct estimation with additional ideas about how to handle identification problems in this setting. The parameter constraints we chose borrow from Bayesian ideas because they effectively shrink what would otherwise be extreme estimates to be less extreme. The approach performed reasonably well in our simulation studies and so may be worth considering in applications. The R package accompanying this article includes a function for direct estimation that implements this approach. We also conducted simulations that evaluated adding a flattening constant of 0.5 to all counts for any group with sparse counts to enforce existence of the MLE (Fienberg & Holland, 1972), but this resulted in direct estimators with notably worse bias and variance than those obtained by constraining parameters for groups with sparse data. There are also other approaches for dealing with existence problems of MLEs that could be considered in this setting (e.g., Firth, 1993; Warm, 1989), and evaluating the costs and benefits of these possibilities relative to alternatives would be worth future study.

There are several other areas of future work to consider. While our simulations varied the group sample sizes, the number of groups, and the cut point locations, there are numerous other relevant features that could be manipulated and future work could consider the implications of these factors for performance of various estimators. Examples include the incorporation of multiple covariates, different distributions of the true group parameters, violations of the normality assumption of the within-group achievement distributions, the extension to variable group sizes, and the consideration of target estimands beyond those considered here. It also would be reasonable to consider the performance of various estimators using coarsened, aggregate data to corresponding estimators based on individual-level data to better understand the costs of operating with the coarsened, aggregate data. Also, the FH-HETOP modeling framework could be generalized to relax the assumption of normality of the latent variables within groups. A straightforward extension would specify the within-group distributions to be members of the three-parameter skew normal family (Azzalini & Dalla Valle, 1996). In that case, the residual distribution would be three-dimensional, and the conditional regression method of chaining Efron priors would carry over. Regarding the Efron priors, further investigation of the implementation details is warranted. Efron (2016) provides limited guidance on the selection of p, the range of the grid, or specification of $\mathbf{Q}$. Our empirical example with the ECLS-K data illustrated the use of WAIC to select p, but we did not consider more nuanced issues such as how to choose the knot locations determining the elements of $\mathbf{Q}$. The oscillating behavior of the WAIC for small values of p suggests that the alignment of knot locations with the underlying distribution of latent variables can consequentially affect some indicators of fit, and it seems likely that better fit could have been achieved with alternative methods for computing $\mathbf{Q}$. More generally, our empirical example considered only WAIC for model selection with the FH-HETOP model, and it would be reasonable for future work to consider alternative approaches to model selection such as cross validation. Finally, a shortcoming of the TG method of

group parameter estimation is that the algorithmic definition of the TG estimators does not lend itself to an obvious uncertainty measure. The posterior variance of the corresponding parameters may be a reasonable approximation, but future work is needed to evaluate this and potential alternatives.

Acknowledgments

We thank Li Cai, Shelby Haberman, Daniel F. McCaffrey, Sandip Sinharay, and two anonymous referees for their helpful comments on earlier drafts.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: J. R. Lockwood and K. E. Castellano prepared the work as employees of Educational Testing Service. The opinions expressed are those of the authors and do not represent views of ETS, the Institute of Education Sciences, or the U.S. Department of Education.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140032 to ETS.

Notes

1. As described in a later section, the computation of the direct estimates by maximum likelihood estimation (MLE) implicitly uses data from all groups to estimate "cut points" that influence performance category probabilities, but conditional on the estimated cut points, the parameter estimates for a given group are determined by only its own count data.
2. Intuitively, this is because when $K = 3$, each group has two free probabilities to inform the two unknown group parameters conditional on the two fixed cut points. Alternatively, when $K > 3$, each group has $K - 1$ free probabilities to inform the two unknown group parameters and the locations of the $K - 3$ free cut points, again conditional on the two fixed cut points.
3. We follow the suggestion of Reardon et al. (2017) by applying a bias correction to obtain estimates of $\{\mu_g^*, \sigma_g^*\}_{g=1}^G$ from MLEs computed for $\{\mu_g, \sigma_g\}_{g=1}^G$ under alternative identification constraints. Thus, $\{\widehat{\mu}_g^*, \widehat{\sigma}_g^*\}_{g=1}^G$ are technically bias-corrected MLEs rather than true MLEs. However, the substantive results were the same if we instead examined true MLEs.
4. We fit all the considered models with $M = 50$ and found comparable posterior means and intervals for all the regression coefficients, indicating that if computational burden is a concern, a smaller value of M could be used without risk of key inferences changing.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679. doi:10.1080/01621459.1993.10476321

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715–726. doi:10.1093/biomet/83.4.715

Carlin, B., & Louis, T. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC Press. doi:10.1201/9781420057669

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*, 83–87. doi:10.2307/2682801

Devine, O., & Louis, T. (1994). A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine*, *13*, 1119–1133. doi:10.1002/sim.4780131104

DeYoreo, M., & Kottas, A. (2017). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, *27*, 71–84. doi:10.1080/10618600.2017.1316280

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, *29*, 285–301. doi:10.1214/13-sts455

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika*, *103*, 1–20. doi:10.1093/biomet/asv068

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *Journal of the American Statistical Association*, *68*, 117–130. doi:10.2307/2284155

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89–102.

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009-2015. *Educational Researcher*, *47*, 221–234. doi:10.3102/0013189X18759524

Fay, R., & Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269–277. doi:10.2307/2286322

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230. doi:10.1214/aos/1176342360

Fienberg, S. E., & Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis*, *2*, 127–134. doi:10.1016/0047-259x(72)90014-0

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38. doi:10.2307/2336755

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. London, England: Chapman & Hall.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. doi:10.1214/ss/1177011136

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, *87*, 533–540. doi:10.2307/2290287

Ghosh, M., & Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, *9*, 55–76. doi:10.1214/ss/1177010647

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. London, England: Chapman & Hall.

Gill, J., & Casella, G. (2009). Nonparametric priors for ordinal Bayesian social science models: Specification and estimation. *Journal of the American Statistical Association*, *104*, 453–454. doi:10.1198/jasa.2009.0039

Gu, Y., Fiebig, D. G., Cripps, E., & Kohn, R. (2009). Bayesian estimation of a random effects heteroscedastic probit model. *The Econometrics Journal*, *12*, 324–339. doi:10.1111/j.1368-423x.2009.00283.x

Haberman, S. J. (1980). Discussion of "regression models for ordinal data." *Journal of the Royal Statistical Society Series B*, *42*, 136–137.

Haberman, S. J. (2005). *Identifiability of parameters in item response models with unconstrained ability distributions* (ETS Research Rep. No. RR-05-24). Princeton, NJ: ETS.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*, 461–465. doi:10.2307/1913974

Hedeker, D., Demirtas, H., & Mermelstein, R. J. (2009). A mixed ordinal location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Statistics and its Interface*, *2*, 391–401. doi:10.4310/sii.2009.v2.n4.a1

Johnson, V. E. (1996). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, *91*, 42–51. doi:10.2307/2291381

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer-Verlag.

Kapur, K., Li, X., Blood, E. A., & Hedeker, D. (2015). Bayesian mixed-effects location and scale models for multivariate longitudinal outcomes: An application to Ecological Momentary Assessment data. *Statistics in Medicine*, *34*, 630–651. doi:10.1002/sim.6345

Kim, J., & Choi, K. (2008). Closing the gap: Modeling within-school variance heterogeneity in school effect studies. *Asia Pacific Education Review*, *9*, 206–220. doi:10.1007/bf03026500

Kubokawa, T., & Strawderman, W. E. (2013). Dominance properties of constrained Bayes and empirical Bayes estimators. *Bernoulli*, *19*, 2200–2221. doi:10.3150/12-bej449

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 215–232. doi:10.2307/2286284

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance-covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*, 307–332. doi:10.3102/1076998614546494

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*, 22–52. doi:10.3102/1076998613509405

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *Annals of Applied Statistics*, *9*, 1484–1509. doi:10.1214/15-aoas833

Louis, T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, *79*, 393–398. doi:10.2307/2288281

Lyles, R. H., Moore, R., Manatunga, A., & Easley, K. (2009). Covariate-adjusted constrained Bayes predictions of random intercepts and slopes. *Journal of Modern Applied Statistical Methods*, *8*, 81–94. doi:10.22237/jmasm/1241136360

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B*, *42*, 109–142.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381. doi:10.1007/bf02306026

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161. doi:10.1111/j.1745-3984.1992.tb00371.x

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*, 343–369. doi:10.1177/0013164413499344

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association*, *78*, 47–55. doi:10.2307/2287098

National Center for Education Statistics. (2002). *User's manual for the ECLS-K first grade public-use data files and electronic code book* (NCES 2002-135). Retrieved from https://nces.ed.gov/pubs2002/2002135_2.pdf

Ohlssen, D. I., Sharples, L. D., & Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, *26*, 2088–2112. doi:10.1002/sim.2666

Paddock, S. M., & Louis, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*, 575–589. doi:10.1111/j.1467-9876.2010.00760.x

Paddock, S. M., Ridgeway, G., Lin, R., & Louis, T. A. (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics & Data Analysis*, *50*, 3243–3262. doi:10.1016/j.csda.2005.05.008

Pfefferman, D. (2002). Small area estimation—New developments and directions. *International Statistical Review*, *70*, 125–143. doi:10.2307/1403729

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, *3*, 215–232. doi:10.1191/1471082x03st056oa

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata, Volume II: Categorical responses, counts, and survival* (3rd ed.). College Station, TX: Stata Press.

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., & DiSalvo, R. (2017). *Stanford education data archive* (Version 2.0). Retrieved from http://purl.stanford.edu/db586ns4974

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*, 3–45. doi:10.3102/1076998616666279

Roeder, K., Carroll, R., & Lindsay, B. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, *91*, 722–732. doi:10.2307/2291667

Savitsky, T. D., & McCaffrey, D. F. (2014). Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika*, *79*, 275–302. doi:10.1007/s11336-013-9339-z

Segawa, E. (2005). A growth model for multilevel ordinal data. *Journal of Educational and Behavioral Statistics*, *30*, 369–396. doi:10.3102/1076998603000 4369

Shen, W., & Louis, T. (1998). Triple-goal estimates in two-stage, hierarchical models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *60*, 455–471. doi:10.1111/1467-9868.00135

Shen, W., & Louis, T. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, *8*, 800–823. doi: 10.2307/1390828

Shen, W., & Louis, T. (2000). Triple-goal estimates for disease mapping. *Statistics in Medicine*, *19*, 2295–2308.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *64*, 583–639. doi:10.1111/1467-9868. 00353

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi:10.1007/s11222-016-9696-4

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/bf02294627

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*, 281–301. doi: 10.1007/s11336-004-1175-8

Authors

J. R. LOCKWOOD is a principal research scientist at ETS, 660 Rosedale Road, Princeton, NJ 08541, USA; email: jrlockwood@ets.org. His research interests include measurement error modeling and causal inference with observational data.

KATHERINE E. CASTELLANO is a research scientist at ETS, 660 Rosedale Road, Princeton, NJ 08541, USA; email: kecastellano@ets.org. Her research interests involve the application of statistical models to addressing educational policy issues, such as the use of student growth models in accountability systems.

BENJAMIN R. SHEAR is an assistant professor in the Research and Evaluation Methodology program at the University of Colorado Boulder's School of Education,

249 UCB, Boulder, CO 80309, USA; email: benjamin.shear@colorado.edu. His research focuses on applied statistical issues in educational measurement and psychometrics, particularly those relevant for validity and validation.

Manuscript received October 3, 2017
Revision received June 5, 2018
Accepted June 17, 2018