Using Auxiliary Item Information in the Item Parameter Estimation of a Graded Response Model for a Small to Medium Sample Size: Empirical versus Hierarchical Bayes Estimation

By

Matthew Naveiras

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER'S OF SCIENCE

in

Psychology

May 31, 2020

Nashville, Tennessee

Approved:

Sun-Joo Cho, Ph.D.

Hao Wu, Ph.D.

Kris Preacher, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Item response theory (IRT) is a test theory that analyzes responses at the item level. IRT is a popular methodology for developing and evaluating scales used in educational and psychological research. In IRT, marginal maximum likelihood estimation (MMLE; Bock & Aitkin, 1981), or full information maximum likelihood (FIML), is largely considered the "gold standard" for item parameter estimation (Baker & Kim, 2004). Previous research on MMLE has shown that the accuracy and precision of item parameter estimates is acceptable in medium and large sample sizes (e.g., Forero & Maydeu-Olivares, 2009; Kieftenbeld & Natesan, 2012; Reise & Yu, 1990). In small sample sizes MMLE can struggle with obtaining accurate and precise item parameter estimates, or may not converge at all. Unfortunately, it is not uncommon for researchers to struggle with obtaining medium or large sample sizes. Studies of rare populations (e.g., individuals with Rett syndrome, students with listening fatigue, individuals with substance use disorders) can make it difficult to obtain more participants. In addition, small research institutions may not have the necessary funds to afford to test more participants. With small sample sizes, alternative methods are required to obtain accurate and precise item parameter estimates with whatever data are available.

Bayesian estimation methods have been used in IRT estimation to increase the accuracy (by reducing the mean squared error) and precision (by reducing the standard error) of item parameter estimates (e.g., Albert, 1992; Edwards, 2010; Patz & Junker, 1999). Mislevy (1988) proposed an empirical Bayes method to increase the stability and precision of item location (or difficulty) estimates in Rasch models. The method proposed by Mislevy (1988) is considered "empirical Bayes" because it uses both maximum likelihood estimates and regression estimates (as prior means) to obtain shrinkage estimates in three steps. However, the implementation of this three-step empirical Bayes method differs from the one-step implementation of empirical Bayes most commonly performed in the literature. We discuss these differences in the summary and discussion section. Mislevy (1988) used auxiliary item information (i.e., item domain information such as what mathematical operation(s) was/were

required to solve items) to compensate for the lack of information available from persons in a sample size of 150. Auxiliary item information was used by the empirical Bayes method as item covariates grouping similar items together regarding their content, format, or the skills required to solve them. In Mislevy's (1988) study, using auxiliary item information resulted in a 25% increase in the precision of item location estimates, an increase that otherwise would have required testing approximately 40 additional persons.

One limitation of the empirical Bayes method used by Mislevy (1988) is that the uncertainty of item parameter estimates is ignored, which can result in underestimated standard errors. This underestimation of standard errors is especially problematic with small sample sizes. To incorporate the uncertainty of item parameter estimates, hierarchical Bayes methods can be used. As opposed to empirical Bayes, which uses point priors for item parameters, hierarchical Bayes methods specify prior distributions on item parameters (called "hyper-priors"). Inverse-gamma(ϵ,ϵ) distributions are typically selected as hyper-priors on the variance of item parameters for their conditional conjugacy (having prior and conditional posterior distributions belonging to the same class) suggesting clean mathematical properties. However, Gelman (2006) does not recommend inverse-gamma(ϵ,ϵ) distributions as noninformative priors, because the resulting inferences when estimating near-zero standard deviations are highly dependent upon the choice of ϵ. Instead, Gelman (2006) recommends half-t distributions (specifically half-Cauchy when the number of groups is small) on standard deviations as weakly-informative and conditionally-conjugate priors.

Likert-type rating scales are common in psychological research. The item response model most widely used for modeling rating scales is the graded response model (GRM; Samejima 1969). The GRM is popular for being highly flexible in modeling tests with items having unique thresholds (both in number and in location for each item). No previous research has been conducted on applying the empirical Bayes method (as used by Mislevy, 1988) to the GRM or on evaluating the performance of using half-t and half-Cauchy distributions in a hierarchical Bayes method as hyper-priors for the GRM.

The primary purpose of this study is to apply empirical and hierarchical Bayes methods using auxiliary item information to a GRM to obtain item parameter estimates with greater accuracy and precision, particularly in small to medium sample sizes. For the purpose of comparing empirical Bayes and hierarchical Bayes, we extend Mislevy's (1988) empirical Bayes method for a Rasch model to a GRM. Specific research questions this study plans to answer regarding the GRM are as follows:

(1a) Among the estimation methods of interest (MMLE, empirical Bayes, and hierarchical Bayes), which method results in the most accurate item parameter estimates in small to medium sample sizes?

(1b) Is a hierarchical Bayes method an acceptable alternative to MMLE in small to medium sample sizes when MMLE is unable to achieve convergence?

(2) How much is the accuracy of item parameter estimates in small to medium sample sizes increased when using a hierarchical Bayes method with the use of item covariates compared to a hierarchical Bayes method without item covariates?

(3) How much is the underestimation of the standard errors of item parameter estimates reduced in small to medium sample sizes by including the uncertainty of item parameter estimates with a hierarchical Bayes method compared to an empirical Bayes method?

These research questions will be answered by comparing the results of MMLE, empirical Bayes, and hierarchical Bayes (with and without the use of item covariates) via a simulation study. An additional research goal of this study is to provide `R` functions for the application of these empirical and hierarchical Bayes methods.

The rest of this paper is structured as follows. First, a background of the GRM with auxiliary item information and of Bayesian analysis and its relevant concepts are provided. Second, empirical and hierarchical Bayes methods are described. Third, the methods described are applied to an empirical data set to illustrate MMLE, empirical Bayes, and hierarchical Bayes methods. Fourth, a simulation study is conducted to evaluate the relative performance of the methods described under various simulation conditions. Finally, we conclude with a

summary and discussion.

Background

GRM with Auxiliary Item Information

Samejima's (1969) GRM specifies the conditional cumulative probability of response y_{ji} for person j ($j = 1, \ldots, J$) and item i ($i = 1, \ldots, I$) in category k ($k = 0, 1, \ldots, m_i - 1$), where m_i is the number of categories for each item i, as follows:

$$P(y_{ji} \geq k|\theta_j) = \tag{1}$$

$$\begin{cases} 1 & \text{if} \quad k = 0 \\ logit^{-1}[\alpha_i(\theta_j - \beta_{i,k})] & \text{if} \quad 1 \leq k \leq m_i - 1, \end{cases}$$

where $logit^{-1}$ denotes the inverse logit link, α_i is an item discrimination parameter, $\beta_{i,k}$ is an item threshold parameter, and θ_j is a latent variable. The categorical response probability is specified as the difference between two adjacent cumulative probabilities:

$$P(y_{ji} = k|\theta_j) = P(y_{ji} \geq k|\theta_j) - P(y_{ji} \geq k + 1|\theta_j), \tag{2}$$

where $P(y_{ji} \geq m_i|\theta_j) = 0$.

Variability in item parameters across items can be explained or predicted using auxiliary item information such as item format (e.g., Hohensinn & Kubinger, 2011), item contents (or domains, e.g., Shermis & Chang, 1997), or the skills required to solve items (e.g., Hartig, Frey, Nold, & Klieme, 2012). There are three lines of research on the use of auxiliary item information. The first line of research is to use auxiliary item information to obtain stable and precise item parameter estimates of item response models. Mislevy (1988) presented an empirical Bayes method to obtain stable and precise Rasch item location parameter estimates. The second line of research is to incorporate cognitive theories about the skills required to answer an item correctly in item response models (Embretson, 1998). The third line of research is to use auxiliary item information for item generation models (Bejar, 2012; Cho, De Boeck, Embretson, & Rabe-Hesketh, 2014; Embretson, 1999).

In this paper, we focus on the use of auxiliary item information to obtain stable and precise item parameter estimates of the GRM using empirical and hierarchical Bayes methods. A linear regression model with normal and homoscedastic residuals is assumed for item parameters, as used in other item regression models (e.g., Cho et al., 2014; Mislevy, 1988). The regression structure of item discrimination parameters can be imposed as follows:

$$\alpha_i = \gamma_{\alpha 0} + \sum_{d=1}^{D} \gamma_{\alpha d} x_{id} + \epsilon_{\alpha i}, \tag{3}$$

where d is the index for auxiliary item information (or item covariate) ($d = 1, \ldots, D$), $\gamma_{\alpha 0}$ is the intercept parameter, $\gamma_{\alpha d}$ is the effect of item covariate x_{id} on discrimination parameter α_i, and $\epsilon_{\alpha i}$ is the random item residual, assumed to follow $(\epsilon_{\alpha 1}, \ldots, \epsilon_{\alpha I})^T \sim MVN(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_I)$, where σ_α^2 is the variance of the random item residual. Similarly, for item threshold parameters:

$$\beta_{i,k} = \gamma_{\beta 0 k} + \sum_{d=1}^{D} \gamma_{\beta d k} x_{id} + \epsilon_{\beta i k}, \tag{4}$$

where $\gamma_{\beta 0 k}$ is the intercept parameter, $\gamma_{\beta d k}$ is the effect of item covariate x_{id} on threshold parameter $\beta_{i,k}$, and $\epsilon_{\beta i k}$ is the random item residual, assumed to follow $(\epsilon_{\beta 1 k}, \ldots, \epsilon_{\beta I k})^T \sim MVN(\mathbf{0}, \sigma_{\beta k}^2 \mathbf{I}_I)$, where $\sigma_{\beta k}^2$ is the variance of the random item residual across items for category k.

Bayesian Analysis

Bayesian analysis uses prior distributions on a set of parameters S (denoted by $P(S)$) and the likelihood of available data Y (denoted by $P(Y|S)$) to create a posterior distribution (denoted by $P(S|Y)$) that generally results in more stable estimates with smaller standard errors than maximum likelihood estimates (e.g., Gelman et al., 2014). The important concepts of Bayesian analysis and their applications to the GRM as used in this paper are described below.

Prior distribution. In Bayesian analysis, the prior distributions are the probability distributions on the set of parameters S before using the data. This distribution can come from past research, or from a researcher's "best guess" of what the population's distribution

looks like. In the GRM, the set of parameters is denoted by $S = \{\theta_j, \alpha_i, \beta_{ik} \mid (1 \leq j \leq J), (1 \leq i \leq I), (1 \leq k \leq m_i - 1)\}$, which includes the level of the latent trait for each person (θ_j), as well as the item discrimination and threshold parameters for each item (α_i and β_{ik}, respectively). The prior distributions of these parameters are denoted by $P(\theta_j)$, $P(\alpha_i)$, and $P(\beta_{ik})$, respectively, with an independent prior assumption.

Exchangeability. When we make the assumption of exchangeability we assume that parameters of the same type are obtained from the same population at random, and as a result the ordering of their subscripts is purely arbitrary and has no bearing on their values. Under this assumption, the joint probability distribution of parameters is invariant to the permutation of its subscripts.

For the GRM, we make the assumption that parameters are identically and independently distributed (iid), meaning that parameters of the same type (such as levels of the latent trait over persons, discrimination parameters over items, and threshold parameters within a category over items) are obtained from the same population for that parameter's type independently of one another, and therefore the value of any parameter has no effect on the value of any other parameter of the same type.

For the GRM, the iid sequence is exchangeable. Therefore, exchangeability is assumed for the GRM for levels of the latent trait (θ_j) over persons. As an example, $P(\theta_1, \theta_2, \theta_3) = P(\theta_3, \theta_1, \theta_2)$ can be assumed for persons $j = 1, 2, 3$. Similarly, discrimination parameters are considered to be exchangeable over items, as are threshold parameters (within a category) over items. Assuming exchangeability over items can be justified when there is no prior information to distinguish among the items. However, when auxiliary item information is available, the exchangeability of items can only be assumed for items with identical auxiliary item information.

Likelihood. Once the data are obtained, the likelihood function $P(Y|S)$ can be used to specify the probability that any possible set of parameters S could result in the data obtained. The higher the likelihood for a set of parameters S, the more likely S is to reflect

the population from which the sample was obtained. Given the assumptions of iid for the GRM, $P(Y|S)$ can be obtained by taking the product of all likelihood functions for individual responses y_{ji} as follows:

$$P(Y|S) = \prod_{j=1}^{J} \prod_{i=1}^{I} \prod_{k=0}^{m_i-1} P(y_{ji} \geq k \mid S)^{I[y_{ji}=k]}, \tag{5}$$

where $I[y_{ji} = k]$ is the indicator function, equal to 1 when $y_{ji} = k$ and otherwise equal to 0. **Posterior distribution using Bayes' theorem.** Once the likelihood function is obtained, the posterior distribution $P(S|Y)$ can be specified using Bayes' theorem:

$$P(S|Y) = \frac{P(Y|S) \cdot P(S)}{\int_S P(Y|S) \cdot P(S) dS} \propto P(Y|S) \cdot P(S), \tag{6}$$

where $P(Y|S)$ is the likelihood function, $P(S)$ is the prior distribution of S, and $\int_S P(Y|S) \cdot P(S)dS$ is equivalent to the numerator of Bayes' theorem integrated over all possible parameter sets S. Because this integral is fixed, the posterior distribution $P(S|Y)$ is proportional to the numerator of Equation 6, $P(Y|S) \cdot P(S)$.

Given the assumptions of iid and exchangeability for the GRM, the joint posterior distribution for S is specified as follows:

$$P(S|Y) \propto P(Y|S) \cdot \prod_{j=1}^{J} \prod_{i=1}^{I} \prod_{k=1}^{m_i-1} P(\theta_j) P(\alpha_i) P(\beta_{ik}), \tag{7}$$

where the joint probability $\prod_{j=1}^{J} \prod_{i=1}^{I} \prod_{k=1}^{m_i-1} P(\theta_j)P(\alpha_i)P(\beta_{ik})$ is equal to $P(S)$.

Empirical vs. hierarchical Bayes method. If it is assumed that parameters of the same type (such as latent trait θ_j) are obtained from the same population (e.g., $N(\mu_\theta, \sigma_\theta^2)$), then the parameters of that population (called hyperparameters, denoted by $H = \{\mu_\theta, \sigma_\theta^2\}$) need to be estimated as well. There are two different approaches for estimating H.

The first approach, called empirical Bayes, obtains estimates for H by working backwards from the data to determine what values of H are most likely to result in the data obtained. Although this method is the simpler of the two, it is not ideal because it uses the data to

obtain the prior's hyperparameters, which goes against the idea of a prior distribution being obtained before the data are obtained.

The second approach, called hierarchical Bayes, treats H as a set of parameters to be estimated alongside S. Using this approach, Bayes' theorem is rewritten to include H as follows:

$$P(S, H \mid Y) \propto P(Y \mid S, H) \cdot P(S \mid H) \cdot P(H), \tag{8}$$

where $P(S \mid H)$ is the prior distribution of S given hyperparameters H, and $P(H)$ is the hyper-prior distribution. Hierarchical Bayesian analysis is usually preferred over empirical Bayesian analysis for its higher precision in estimating H. Hierarchical Bayesian analysis incorporates the uncertainty in estimating H by including its hyper-prior distribution $P(H)$ in the model. Empirical Bayesian analysis, however, ignores this uncertainty by only obtaining point estimates for H, often resulting in an underestimation of the posterior standard deviations for item parameter estimates.

Shrinkage estimator. In the prior distribution, all parameters of the same type are restricted to the prior mean. The less variation there is within the population, the closer posterior estimates are to the prior mean. Figure 1 (top) (inspired by Figure 1 in Mislevy [1988]) illustrates this concept for nine different parameter estimates.

In Figure 1 (top), posterior estimates are a compromise between the prior mean and the maximum likelihood estimates (obtained from the data without the use of priors). The posterior estimates, or shrinkage estimates, have higher precision than maximum likelihood estimates. However, these posterior estimates are also subject to larger bias than maximum likelihood estimates. This bias results from the use of the prior mean in obtaining posterior estimates. How much influence the prior mean has over posterior estimates is called "shrinkage," which is calculated as the distance between the maximum likelihood estimate and the posterior estimate relative to the total distance between the maximum likelihood estimate and the prior mean. For example, the proportional shrinkage of item discrimination estimates is calculated as follows:

$$Shrinkage(\tilde{\alpha}_i) = \frac{\widehat{\alpha}_i - \tilde{\alpha}_i}{\widehat{\alpha}_i - \bar{\alpha}_i}, \tag{9}$$

where $\widehat{\alpha}_i - \tilde{\alpha}_i$ is the distance between the maximum likelihood estimate ($\widehat{\alpha}_i$) and the posterior estimate ($\tilde{\alpha}_i$), and $\widehat{\alpha}_i - \bar{\alpha}_i$ is the total distance between the maximum likelihood estimate and the prior mean ($\bar{\alpha}_i$). These distances are illustrated in Figure 1 (middle). Less variation within the population (i.e., prior variance) results in posterior estimates shrinking more towards the prior mean. In the extreme case where the prior variance is equal to 0, posterior estimates shrink completely to equal the prior mean. Inversely, more variation within the population results in posterior estimates shrinking less towards the prior mean. In the extreme case where the prior variance is infinite, posterior estimates won't shrink at all, resulting in posterior estimates equaling the maximum likelihood estimates.

Gelman et al. (2014) discusses the advantages of using partially-pooled estimates obtained with a prior mean, as this allows both information regarding the similarities (pooled estimates) and individualities (separate estimates) of the data to be utilized. The usage of both pooled and separate information results in estimates with higher accuracy. When there is no prior information to distinguish item parameters of the same type, we assume exchangeability among all such parameters. When this is the case, all item parameters of the same type shrink towards a single prior mean, as illustrated in Figure 1 (top). When item covariates are available to indicate different item groups, item parameters are considered exchangeable only with item parameters within the same group (i.e., having the same values for item covariates [Mislevy, 1988]). When this is the case, each group of item parameters shrinks towards a different prior mean, as illustrated in Figure 1 (bottom) with three item covariates. As shown in Gelman et al. (2014), shrinkage estimates with different prior means tend to be less biased than shrinkage estimates with a single prior mean.

Methods

In this section we describe the empirical Bayes and hierarchical Bayes methods implemented in this study, and how these methods can be used to obtain estimates of GRM item

parameters by using auxiliary item information. We extend Mislevy (1988)'s empirical Bayes method for the Rasch model to the method for the GRM, and then discuss the specification of the prior and posterior distribution for hierarchical Bayes.

Empirical Bayes Method

The estimation of GRM item parameters with an empirical Bayes method takes place over three steps, as described below.

Step 1. Maximum likelihood estimates of item parameters. In Step 1, item parameters (α_i and $\beta_{i,k}$) and corresponding standard errors ($\tau_{\alpha i}$ and $\tau_{\beta ik}$) were estimated using MMLE. MMLE was implemented using the `mirt` package (Chalmers, 2019) in `R` (R Core Team, 2018).

Step 2. Maximum likelihood estimates of the regression parameters and the residual variance. In Step 2, we consider item regression models (Equations 3 and 4) using the maximum likelihood estimates of item parameters obtained in Step 1 ($\widehat{\alpha}_i$ and $\widehat{\beta}_{ik}$). Because we use the maximum likelihood estimates from Step 1, the uncertainty of these estimates is ignored in Step 2. Maximum likelihood estimates of the regression parameters of these item regression models were obtained using the `lm` function in `R`.

The regression structure is imposed on item discrimination estimates as follows:

$$\widehat{\alpha}_i = \gamma_{\alpha 0} + \Sigma_{d=1}^{D} \gamma_{\alpha d} x_{id} + h_{\alpha i}, \tag{10}$$

where $(h_{\alpha 1}, \ldots, h_{\alpha I})^T \sim MVN(\mathbf{0}, \phi_\alpha^2 I)$. Similarly, for item threshold estimates:

$$\widehat{\beta}_{ik} = \gamma_{\beta 0k} + \Sigma_{d=1}^{D} \gamma_{\beta dk} x_{id} + h_{\beta ik}, \tag{11}$$

where $(h_{\beta 1k}, \ldots, h_{\beta Ik})^T \sim MVN(\mathbf{0}, \phi_{\beta k}^2 I)$.

Unbiased estimates of the residual variances (ϕ_α^2 and $\phi_{\beta k}^2$) were calculated using the following equations (Rencher, 2000, p. 143):

$$\widehat{\phi}_\alpha^2 = \frac{\sum_{i=1}^{I} \tilde{h}_{\alpha i}^2}{I - D - 1} \tag{12}$$

and

$$\widehat{\phi}^2_{\beta k} = \frac{\sum_{i=1}^{I} \tilde{h}^2_{\beta ik}}{I - D - 1}, \tag{13}$$

where D is the number of item covariates. The standard errors of the residual variance for item discrimination estimates were calculated using the following equation (Rencher, 2000, p. 143):

$$\frac{(I - D - 1)\widehat{\phi}^2_{\alpha}}{\phi^2_{\alpha}} \sim \chi^2(I - D - 1). \tag{14}$$

The variance of each side in Equation 14 is:

$$Var[\frac{(I - D - 1)\widehat{\phi}^2_{\alpha}}{\phi^2_{\alpha}}] = 2(I - D - 1). \tag{15}$$

Accordingly, the standard errors of the residual variance for item discriminations can be calculated as follows:

$$SE_{\phi^2_{\alpha}} = \sqrt{Var(\widehat{\phi}^2_{\alpha})} = \sqrt{2(I - D - 1)(\frac{\phi^2_{\alpha}}{I - D - 1})^2} = \sqrt{\frac{2\phi^4_{\alpha}}{I - D - 1}}. \tag{16}$$

Following a similar derivation, the standard errors of the residual variances for item threshold estimates were calculated as follows:

$$SE_{\phi^2_{\beta k}} = \sqrt{\frac{2\phi^4_{\beta k}}{I - D - 1}}. \tag{17}$$

Step 3. Empirical Bayes estimates of item parameters. In Step 3, the empirical Bayes estimates of item parameters and the precision of those estimates are calculated, based on the results obtained from Steps 1 and 2. The empirical Bayes estimate $\tilde{\alpha}_i$ is the weighted average of the maximum likelihood estimate $\widehat{\alpha}_i$ and the regression estimate $\bar{\alpha}_i = \widehat{\gamma}_{\alpha 0} + \Sigma_{d=1}^{D} \widehat{\gamma}_{\alpha d} x_{id}$ with weights proportional to their respective precisions[1]:

$$\tilde{\alpha}_i = E(\alpha | \widehat{\alpha}_i, \widehat{\tau}^2_{\alpha i}, \widehat{\gamma}_{\alpha 0}, \widehat{\gamma}_{\alpha d}, \widehat{\phi}^2_{\alpha}) = \frac{\widehat{\alpha}_i \widehat{\tau}^{-2}_{\alpha i} + \bar{\alpha}_i \widehat{\phi}^{-2}_{\alpha}}{\widehat{\tau}^{-2}_{\alpha i} + \widehat{\phi}^{-2}_{\alpha}}. \tag{18}$$

Equation 18 implies the following at the extreme cases:

- $\tilde{\alpha}_i = \bar{\alpha}_i$ if $\widehat{\alpha}_i = \bar{\alpha}_i$ or $\phi^2_{\alpha} = 0$. Having $\phi^2_{\alpha} = 0$ means that $\bar{\alpha}_i$ is infinitely more precise

[1]The precision of an estimate is equal to the inverse of its variance.

than $\widehat{\alpha}_i$.

- $\tilde{\alpha}_i = \widehat{\alpha}_i$ if $\widehat{\alpha}_i = \bar{\alpha}_i$ or $\tau^2_{\alpha i} = 0$. Having $\tau^2_{\alpha i} = 0$ means that $\widehat{\alpha}_i$ is infinitely more precise than $\bar{\alpha}_i$.

Similarly for item threshold parameters, the empirical Bayes estimate $\tilde{\beta}_{ik}$ is the weighted average of the maximum likelihood estimate $\widehat{\beta}_{ik}$ and the regression estimate $\bar{\beta}_{ik} = \widehat{\gamma}_{\beta 0k} + \Sigma^D_{d=1}\widehat{\gamma}_{\beta dk}x_{id}$ with weights proportional to their respective precisions:

$$\tilde{\beta}_{ik} = E(\beta|\widehat{\beta}_{ik}, \widehat{\tau}^2_{\beta ik}, \widehat{\gamma}_{\beta 0k}, \widehat{\gamma}_{\alpha d}, \widehat{\phi}^2_{\beta k}) = \frac{\widehat{\beta}_{ik}\widehat{\tau}^{-2}_{\beta ik} + \bar{\beta}_{ik}\widehat{\phi}^{-2}_{\beta k}}{\widehat{\tau}^{-2}_{\beta ik} + \widehat{\phi}^{-2}_{\beta k}}. \tag{19}$$

Each empirical Bayes estimate ($\tilde{\alpha}_i$, $\tilde{\beta}_{ik}$) gains precision from both the precision of its maximum likelihood estimates ($\widehat{\tau}^{-2}_{\alpha i}$, $\widehat{\tau}^{-2}_{\beta ik}$) obtained in Step 1 and from the precision of its regression estimates ($\widehat{\phi}^{-2}_{\alpha}$, $\widehat{\phi}^{-2}_{\beta k}$) obtained in Step 2:

$$\tilde{\sigma}^{-2}_{\alpha i} = \widehat{\tau}^{-2}_{\alpha i} + \widehat{\phi}^{-2}_{\alpha} \tag{20}$$

and

$$\tilde{\sigma}^{-2}_{\beta ik} = \widehat{\tau}^{-2}_{\beta ik} + \widehat{\phi}^{-2}_{\beta k}. \tag{21}$$

The residual variance for the empirical Bayes estimates of each item parameter type is equal to the inverse of its summed precision:

$$\tilde{\sigma}^2_{\alpha i} = \frac{1}{\widehat{\tau}^{-2}_{\alpha i} + \widehat{\phi}^{-2}_{\alpha}} \tag{22}$$

and

$$\tilde{\sigma}^2_{\beta ik} = \frac{1}{\widehat{\tau}^{-2}_{\beta ik} + \widehat{\phi}^{-2}_{\beta k}}. \tag{23}$$

As shown earlier, proportional shrinkage is calculated as the distance between the maximum likelihood estimate and the posterior estimate relative to the total distance between the maximum likelihood estimate and the prior mean. For item discrimination parameters:

$$Shrinkage(\tilde{\alpha}_i) = \frac{\widehat{\alpha}_i - \tilde{\alpha}_i}{\widehat{\alpha}_i - \bar{\alpha}_i}. \tag{24}$$

Similarly, for item threshold parameters:

$$Shrinkage(\tilde{\beta}_{ik}) = \frac{\widehat{\beta}_{ik} - \tilde{\beta}_{ik}}{\widehat{\beta}_{ik} - \bar{\beta}_{ik}}. \tag{25}$$

Hierarchical Bayes Method

Specifications of prior and posterior distributions. For the GRM with auxiliary item information, the joint posterior distribution of $S = \{\theta_j, \alpha_i, \beta_{ik}, \gamma_{\alpha 0}, \gamma_{\alpha d}, \sigma^2_\alpha, \gamma_{\beta 0k}, \gamma_{\beta dk}, \sigma^2_{\beta k}\}$ can be written as:

$$\mathrm{P}(S|\mathbf{y}) \propto \{\prod_{j=1}^{J}\prod_{i=1}^{I}\prod_{k=0}^{m_i-1} P(y_{ji} \geq k \mid S)^{I(y_{ji}=k)}\}\cdot$$

$$\{\prod_{j=1}^{J} P(\theta_j)\}\{\prod_{i=1}^{I} P(\alpha_i|\gamma_{\alpha 0}, \boldsymbol{\gamma}_\alpha, \sigma^2_\alpha)\}\{\prod_{i=1}^{I}\prod_{k=1}^{m_i-1} P(\beta_{ik}|\gamma_{\beta 0k}, \boldsymbol{\gamma}_{\beta k}, \sigma^2_{\beta k})\}\cdot$$

$$P(\gamma_{\alpha 0})P(\gamma_{\alpha d})P(\sigma^2_\alpha)\prod_{k=1}^{m_i-1} P(\gamma_{\beta 0k})P(\gamma_{\beta dk})P(\sigma^2_{\beta k}), \tag{26}$$

where the first line is the likelihood function, the second line is the prior distributions, and the third line is the hyper-prior distributions.

Independent priors for θ_j, α_i, and β_{ik} were specified as follows:

$$\theta_j \sim N(0, 1),$$

$$\alpha_i \sim N(\gamma_{\alpha 0} + \sum_{d=1}^{D}\gamma_{\alpha d}x_{id}, \sigma^2_\alpha),$$

and

$$\beta_{i,k} \sim N(\gamma_{\beta 0k} + \sum_{d=1}^{D}\gamma_{\beta dk}x_{id}, \sigma^2_{\beta k}).$$

We could impose a univariate prior or a multivariate prior on item thresholds in a hierarchical Bayes method. When we use a weakly-informative prior as in the present study, we found that the choice of prior between the univariate prior and the multivariate prior does not affect posterior distributions of item parameters except in the case of highly-correlated item thresholds (e.g., $r = .7$). The detailed investigation is provided in Appendix A.

The hyper-prior distributions on regression coefficients ($\gamma_{\alpha 0}, \gamma_{\alpha d}, \gamma_{\beta 0 k}$, and $\gamma_{\beta d k}$) were set as a normal distribution with weakly informative priors, $N(0, 10)$. Weakly informative priors should be selected to intentionally convey less prior information than is readily available, to eliminate or discourage impossible or improbable parameter values without influencing the posterior in one particular direction over another (Gelman et al., 2014). The weakly informative prior $N(0, 10)$ on regression coefficients (as illustrated in Figure 2 [top]) was selected to indicate a minimal preference towards zero, as these values are generally expected to be relatively small in magnitude.

Gelman (2006) recommended the half-t or half-Cauchy distribution on standard deviation parameters as a weakly-informative and conditionally-conjugative prior, especially when dealing with small sample sizes. Polson and Scott (2011) noted that "the half-Cauchy occupies a sensible 'middle ground' ... it performs very well near the origin, but does not lead to drastic compromises in other parts of the parameter space." The half-Cauchy distribution with a scale parameter of 10 was used on residual SD (RSD) parameters in this study:

$$\sigma_\alpha \sim Cauchy(0, 10)I(0,)$$

and

$$\sigma_{\beta k} \sim Cauchy(0, 10)I(0,),$$

where $I(0,)$ indicates that the distribution is truncated at 0. As shown in Figure 2 (bottom), the distribution becomes a uniform prior density on standard deviations when the scale parameter of the half-Cauchy increases from 1 to 25. The scale parameter of 10 that we chose is considered weakly informative because it has a gentle slope in the tail and allows the data to dominate when the likelihood is strong in the tail.

For hierarchical Bayes estimates, posterior shrinkage can be calculated for each item parameter type based on the proportional reduction in variance:

$$s = \frac{\sigma^2_{prior} - \sigma^2_{posterior}}{\sigma^2_{prior}} \tag{27}$$

Posterior shrinkage near zero indicates that the data provided little information beyond that present in the selected prior, whereas posterior shrinkage near one indicates that the data provided enough information to strongly influence posterior estimates.

Bayesian computation. MCMC sampling was conducted using `rStan`, the `R` interface to Stan (Stan Development Team, 2018). `rStan` is capable of implementing Euclidean Hamiltonian Monte Carlo (HMC; Duane, Kennedy, Pendleton, and Roweth 1987; Neal 1994, 2011), and by default uses the no-U-turn sampler (NUTS; Hoffman & Gelman 2014; Betancourt 2016) extension. NUTS chooses the number of leapfrog steps automatically for each iteration to eliminate user-required input and maximize efficiency. HMC (both basic and NUTS) allows the implementation of unit, diagonal, and dense mass matrices, and both use gradient information from the log probability density to generate systematic motion through the posterior, reducing redundancies in the space explored to decrease autocorrelations between transitions.

Constraints were imposed on several parameters sampled in `rStan` to prevent highly improbable or impossible item parameter values. Item discrimination parameters and residual SDs were constrained to be strictly non-negative ($\alpha_i \geq 0$, $\sigma_\alpha \geq 0$, $\sigma_{\beta k} \geq 0$), and item thresholds were constrained to be in increasing order ($\beta_{i,1} < \beta_{i,2} < \beta_{i,3} < \beta_{i,4}$, see Appendix B).

Illustration

In this section, we illustrate the empirical and hierarchical Bayes methods described in the previous section by applying them to an empirical data set. `R` functions to implement the methods used below are provided in Appendix B.

Data Description

The data analyzed using the methods described above were collected from the Vanderbilt Fatigue Scale for Adults (VFS-A), which was designed to measure listening-related fatigue.

Preliminary research led to the identification of four domains of listening-related fatigue: cognitive, emotional, physical, and social (Davis, Schlundt, Camarata, Bess, & Hornsby, 2020). Using Mplus Version 8.3 (Muthén & Muthén, 1998-2019), exploratory factor analyses were conducted using polychoric correlations (specifically, limited information robust weighted least square estimation with Oblimin rotation and Oblique type) to extract 1, 2, 3, and 4 factors[2] to explore the number and structure of the factors of the VFS-AHL. In Table 1 these four models were compared with standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI). Based on empirically-supported guidelines a model is considered to fit well if $SRMR < .08$, $RMSEA < .06$, $CFI > .95$, and $TLI > .95$ (Hu & Bentler, 1999; Yu, 2002). The unidimensional model was considered a well-fitting model according to the SRMR, CFI, and TLI. Based on these exploratory factor analyses, we considered listening-related fatigue to be a unidimensional construct.

The research version of the VFS-A was analyzed, having 10 five-point Likert-scale items for each of the four domains of listening-related fatigue, for a total of 40 items. A total of 273 participants completed all 40 items. Of these 273 participants, 150 participants were randomly sampled to illustrate the empirical and hierarchical Bayes methods for a small sample size.

Analysis

The four domains of listening-related fatigue items were treated as item covariates for analysis. For dummy variable coding, the social domain was chosen (arbitrarily) as the reference category. The regression structure for item parameters was structured as follows:

$$\alpha_i = \gamma_{\alpha 0} + \gamma_{\alpha 1} x_{i1} + \gamma_{\alpha 2} x_{i2} + \gamma_{\alpha 3} x_{i3} + \epsilon_{\alpha i} \tag{28}$$

and

$$\beta_{i,k} = \gamma_{\beta 0 k} + \gamma_{\beta 1 k} x_{i1} + \gamma_{\beta 2 k} x_{i2} + \gamma_{\beta 3 k} x_{i3} + \epsilon_{\beta i k}, \tag{29}$$

[2]Quartimin, Geomin, and Target rotation methods resulted in similar patterns of factor loadings across all methods.

where $x_{i1} = 1$ for cognitive items, $x_{i2} = 1$ for emotional items, and $x_{i3} = 1$ for physical items. Estimates for the regression coefficients ($\gamma_{\alpha 0}$, $\gamma_{\alpha d}$, $\gamma_{\beta 0k}$, and $\gamma_{\beta dk}$) were obtained from the `lm` function in `R`, using maximum likelihood estimates obtained from `mirt`. The residual variances (ϕ_α and $\phi_{\beta k}$) were calculated using Equations 12 and 13, and standard errors of the residual variances were calculated using Equations 16 and 17, with $I = 40$ items and $D = 3$ item covariates.

The regression structures for hierarchical Bayesian priors were structured as follows:

$$\alpha_i \sim N(\gamma_{\alpha 0} + \gamma_{\alpha 1} x_{i1} + \gamma_{\alpha 2} x_{i2} + \gamma_{\alpha 3} x_{i3}, \sigma^2_\alpha) \tag{30}$$

and

$$\beta_{i,k} \sim N(\gamma_{\beta 0k} + \gamma_{\beta 1k} x_{i1} + \gamma_{\beta 2k} x_{i2} + \gamma_{\beta 3k} x_{i3}, \sigma^2_{\beta k}). \tag{31}$$

For regression coefficients ($\gamma_{\alpha 0}$, $\gamma_{\alpha d}$, $\gamma_{\beta 0k}$, and $\gamma_{\beta dk}$), a non-informative hyper-prior distribution of $N(0, 10)$ was chosen. A weakly-informative hyper-prior distribution of $Cauchy(0, 10)I(0,)$ was imposed on the standard deviations of residuals (σ_α and $\sigma_{\beta k}$).

Analysis was conducted for sample sizes of 150 and 273 to compare shrinkage and increases in precision for small and medium sample sizes. The default arguments for `rStan` of 4 chains, $2,000$ iterations, 1,000 warmup (i.e., burn-in) iterations per chain, and thinning = 1 were used for analysis. Convergence amongst the 4 chains was evaluated using the Gelman-Rubin statistic (Gelman & Rubin, 1992). Note that these arguments were sufficient to achieve sufficient convergence in both sample sizes, having Gelman-Rubin statistics in the range of 0.95 to 1.05 for all parameters. Obtaining results for sample sizes of 150 and 273 in `R` required approximately 1.2 and 2.0 hours (respectively) on a computer with a 2.30GHz processor and 8.00gb of RAM.

Results

The results obtained for $J = 150$ are presented in this section. The results for $J = 273$ are provided in Appendix C. Comparisons of regression coefficient estimates for both empirical Bayes and hierarchical Bayes methods are illustrated in Table 2. Median hierarchical Bayes

estimates were used for calculating hierarchical shrinkage and standardized differences, as well as for comparing hierarchical and empirical Bayes methods.

Results were highly comparable between empirical and hierarchical Bayes methods: $r(18) = .995$, p-value $< .01$ for item regression parameters, and $r(3) = .975$, p-value $< .01$ for the standard deviations of residuals ($\widehat{\phi}_{\alpha}$ and $\widehat{\phi}_{\beta k}$). However, the standard deviation of residuals were generally larger for the empirical Bayes method than for the hierarchical Bayes method.

Table 3 reports the results for empirical and hierarchical Bayes estimates of α_i, and Table 4 reports the results for empirical and hierarchical Bayes estimates of β_{i4} for illustration. The standard deviations of empirical Bayes estimates were lower than the standard errors of maximum likelihood estimates, because of the added information from item covariates. However, even for a small sample size of J=150, maximum likelihood estimates had significantly lower standard errors than regression estimates (as seen in Table 3 by comparing $\widehat{\tau}_{\alpha i}$ and $\bar{\phi}_{\alpha}$), because the information provided by the data far outweighed the information provided by the item covariates. This is further seen in the calculated values for shrinkage. The average shrinkage for items 2-40 in Table 3 (note that shrinkage was not calculated for item 1) was .159, meaning that on average item covariates contributed 15.9% of the information used in estimating α_i, with the data providing the remaining 84.1% of the information. Similar results were obtained for item threshold parameters. Item threshold parameters (β_{i1}, β_{i2}, β_{i3}, and β_{i4}) had average shrinkages across items of 11.2%, 10.2%, 9.1%, and 12.7% (respectively).

The item parameter estimates obtained using empirical Bayes and hierarchical Bayes (as well as their respective SDs or SEs) were highly similar for all item parameters at each sample size.

Simulation Study

A simulation study was conducted to answer the research questions regarding the empirical and hierarchical Bayes methods described as proposed in this paper's introduction. In this section we describe the design and implementation of this simulation study and discuss the results obtained so as to answer these research questions.

Simulation Factors

In this simulation study, five response categories for each item ($m_i = 5$) was set as a fixed simulation factor, as it is the most commonly used number of response categories in GRM applications (e.g., Forero & Maydeu-Olivares, 2009). Three varying simulation factors were considered that would directly affect item parameter recovery when using the empirical and hierarchical Bayes methods: (a) the number of persons, (b) the number of items, and (c) the RSD of item parameters. Each of these factors is discussed below:

Number of persons. The accuracy of item parameter estimates is mainly affected by the number of persons (Kieftenbeld & Natesan, 2012). Kieftenbeld and Natesan (2012) showed minimal difference in GRM item parameter recovery between MMLE and Markov chain Monte Carlo (MCMC) in sample sizes of 300 or more persons (for 5, 10, 15, and 20 items). Reise and Yu (1990) and Ankermann and Stone (1992) recommended a minimum sample size of 500 to accurately estimate GRM item parameters. Based on this information, sample sizes of 100, 150, 200, 250, 300, and 500 were selected to compare the effectiveness of empirical Bayes and hierarchical Bayes methods at both small sample sizes (100, 150, 200, 250, and 300), and at a medium sample size of 500, which would be considered the maximum sample size at which the empirical Bayes and hierarchical Bayes methods described would be expected to recover item parameters with a performance comparable to MMLE.

Number of items. The number of items affects the accuracy of item covariate effect estimates, as well as the residual variance (e.g., Cho, De Boeck, & Lee, 2017). A literature review we conducted on 28 published papers on the use of item covariates in IRT[3] indicated

[3]Papers published in these six journals were reviewed to report how item covariate structures were used for item response models in common practice: *Acta Psychologica*, *Applied Psychological Measurement (APM)*,

that the number of items ranged from 5 to 334, with a median of 27.5 items (see Table 5). To allow for an equal number of items per item group (to control for the effect of the number of items per item covariate), 24 items were selected for simulation conditions, with each item group having 4 items for 6 item covariates (as explained below). The level of 24 items is close to the common test length of 25 items in the evaluation of GRM parameter recovery (Reise & Yu, 1990). To investigate the effect of test length on item parameter recovery, twice as many items (48) was selected as well, with each item group having 8 items for 6 item covariates (as explained below). Note that this level for the number of items was also selected to reflect the number of items commonly used in large-scale achievement tests (~ 50 items), while still allowing an equal number of items per item group.

RSD of item parameters. The amount of shrinkage is positively affected by the precision of the prior distribution. Note that in Equation 9, shrinkage increases as $\tilde{\alpha}_i \rightarrow \bar{\alpha}_i$, which occurs as a direct result of the precision of $\bar{\alpha}_i$ increasing. Therefore, in order to indirectly manipulate shrinkage in simulation conditions, the RSD of item parameter types (σ_α^2 and $\sigma_{\beta k}^2$) are directly manipulated. Fischer and Rose (2019) considered three levels for the standard deviations of item discrimination and item threshold parameters for GRMs in normal prior distributions: $\sigma_\alpha = \sigma_{\beta k} = 0.5$ (as a weakly informative prior), $\sigma_\alpha = \sigma_{\beta k} = 0.3$ (as a moderately informative prior), and $\sigma_\alpha = \sigma_{\beta k} = 0.1$ (as a strongly informative prior). As the authors noted, the value of an item discrimination, $\alpha_i = 2$ has a 95% probability of being between $1.02 \leq \alpha_i \leq 2.98$ with the weakly informative prior ($2 \pm [1.96 \times 0.5]$) and between $1.804 \leq \alpha_i \leq 2.196$ with the strongly informative prior ($2 \pm [1.96 \times 0.1]$). These same levels of RSD for item discrimination and item threshold parameters were selected for the current study. The weakly informative prior ($\sigma_\alpha = \sigma_{\beta k} = 0.5$) was close to the standard deviation of true item discrimination and item threshold (for the last category) parameters for a GRM that Kieftenbeld and Natesan (2012) and Lautenschlager, Meade, and Kim (2006,

Educational and Psychological Measurement (EPM), *Journal of Educational Measurement (JEM)*, *Multivariate Behavioral Research (MBR)*, and *Psychometrika (PMET)*. Papers were searched using the keywords "item response theory linear logistic test model."

p. 7) used to evaluate item parameter recovery.

The item covariate structure was set as fixed for all simulation conditions because it does not affect accuracy and precision of item parameter estimates directly. The most common item covariate structure (called the Q-matrix) and the number of item covariates selected were based on a literature review of 28 published papers on the use of item covariates to explain item variability in item response models (see Table 5).

Item covariate structure. The item-covariate structure can be specified in a matrix called a Q-matrix. Q-matrices generally took on one of four common patterns. First, 36% (10) of the reviewed papers used a *mutually-exclusive binary Q-matrix* item covariate structure. Items constructed in this way were assigned a value of 1 for at most one of the item covariates, and a value of 0 for all other item covariates. Second, 25% (7) of the reviewed papers used a *non-mutually-exclusive binary Q-matrix* item covariate structure. Items constructed in this way were assigned a value of 1 for any number of item covariates, and a value of 0 for all other item covariates. Third, 11% (3) of the reviewed papers used a *non-mutually-exclusive non-binary Q-matrix* item covariate structure. Items constructed in this way were assigned a value to each item covariate indicative of how many occurrences of that item trait were present in the item. Fourth, 29% (8) of the reviewed papers used a *Q-matrix by factor* item covariate structure. Items constructed in this way had one of the three previously defined item covariate structures for each of two or more item factors, although the most common pattern of this structure was a mutually-exclusive binary Q-matrix for each factor. Because of these observations, a mutually-exclusive binary Q-matrix item covariate structure (the predominant structure observed in the literature) was selected for all simulation conditions.

Number of item covariates. The literature review showed that the number of item groups ranged from 2 to 77, with a median of 6 item groups. Therefore, 6 item covariates were considered. Five dummy-coded item covariates (for five item groups with a sixth reference group) were used.

Effects of item covariates. The effects of item covariates on item discriminations ($\boldsymbol{\gamma}_{\alpha} = [\gamma_{\alpha 0}, \gamma_{\alpha 1}, \gamma_{\alpha 2}, \gamma_{\alpha 3}, \gamma_{\alpha 4}, \gamma_{\alpha 5}]'$) were selected as $[0.813, 0.075, 0.150, 0.225, 0.300, 0.375]'$. The effects of item covariates on item thresholds ($\boldsymbol{\gamma}_{\beta k} = [\gamma_{\beta 0k}, \gamma_{\beta 1k}, \gamma_{\beta 2k}, \gamma_{\beta 3k}, \gamma_{\beta 4k}, \gamma_{\beta 5k}]'$) were set as

$[-2.458, 0.183, 0.367, 0.550, 0.733, 0.917]'$ for the first threshold,

$[-1.458, 0.183, 0.367, 0.550, 0.733, 0.917]'$ for the second threshold,

$[0.542, 0.183, 0.367, 0.550, 0.733, 0.917]'$ for the third threshold, and

$[1.542, 0.183, 0.367, 0.550, 0.733, 0.917]'$ for the fourth threshold. The intercepts of the item thresholds we selected ($\boldsymbol{\gamma}_{\beta 0} = [-2.458, -1.458, 0.542, 1.542]'$) are close to the means of true GRM item thresholds ($[-2.369, -1.334, -0.208, 1.981]'$) that Kieftenbeld and Natesan (2012) and Lautenschlager, Meade, and Kim (2006, p. 7) used in evaluating parameter recovery of GRM item thresholds. The same effects of item covariates were selected for all item thresholds

($[0.183, 0.367, 0.550, 0.733, 0.917]'$) to control for differential item covariate effects across thresholds when investigating item parameter recovery and the precision of item parameter estimates. Given the residual standard deviations of item parameters that we selected, the selected item covariate effects resulted in average adjusted $R^2 = 0.544$ across item parameters when $\sigma_{\alpha} = \sigma_{\beta k} = 0.1$, average adjusted $R^2 = 0.254$ across item parameters when $\sigma_{\alpha} = \sigma_{\beta k} = 0.3$, and average adjusted $R^2 = 0.111$ across item parameters when $\sigma_{\alpha} = \sigma_{\beta k} = 0.5$.

Based on the effects of the item covariates and RSDs described above, true item parameters were calculated during data generation. The latent variable was generated from a standard normal distribution to match it to a model identification constraint. When generating item responses, the same generated item parameters were used across replications and the latent variable was generated for each replication.

The three simulation factors were fully crossed, yielding 36 conditions ($= 6 \times 2 \times 3$). Five hundreds replications were simulated for each of the 36 conditions. Each generated data set

was analyzed using four estimation methods: MMLE, empirical Bayes, hierarchical Bayes with item covariates, and hierarchical Bayes without item covariates.

Convergence Checking

The Gelman and Rubin statistic, $\widehat{R}$, was used to check for convergence when using 4 chains. Five replications of each condition were used to determine the number of warm-up (also known as "burn-in") iterations required for each condition to achieve acceptable convergence ($\widehat{R} < 1.1$, as used in Gelman & Shirley, 2001) for all estimated parameters in these five replications. Across the 36 conditions, the number of warm-up iterations ranged from $1,000$ (for 31 conditions) to $5,000$ (for 1 condition). The number of post-warm-up iterations was set equal to the number of warm-up iterations, resulting in the total number of iterations ranging from $2,000$ to $10,000$.[4] Monte Carlo standard error (MCSE) was examined to evaluate post-burn-in convergence ($MCSE < 0.01 \times SD$).

Evaluation Measures

Three evaluation measures were used to compare the accuracy of the estimates obtained using the four estimation methods (MMLE, empirical Bayes, hierarchical Bayes with item covariates, and hierarchical Bayes without item covariates): absolute relative percentage bias (RPB), root mean square error (RMSE), and absolute relative percentage SD bias (SDB).[5]

To answer research questions 1a and 2, the RPB and RMSE of item parameter estimates are compared between each pair of methods (comparing MMLE, empirical Bayes, and hierarchical Bayes with item covariates in research question 1a, and comparing hierarchical Bayes with and without item covariates in research question 2). To answer research question 1b (regarding the use of hierarchical Bayes as a substitute to MMLE when MMLE fails to

[4]The condition with 300 persons, 48 items, and RSD=0.1 experienced convergence problems for some parameters in 2 out of 5 replications with 10,000 iterations. Thus, to increase the reliability of convergence checking, we checked convergence with 5 additional replications for this condition. Convergence was achieved for 7 of these 10 total replications, and the other three of these replications had relatively acceptable $\widehat{R}$ values, with maximum $\widehat{R}$ values across parameters of 1.581, 1.737, and 1.115. Based on this investigation, we used 10,000 warm-up iterations for this condition.

[5]The absolute values of RPB and SDB are used so that RPB and SDB could be directly comparable among the three methods and five item parameter types, regardless of whether they were positive or negative. The original values for RPB, RMSE, and SDB (non-absolute and separated by parameter type) are provided in Appendix D.

converge), we examine the RPB and the absolute relative percentage differences between posterior SD estimates and the Monte Carlo standard errors (MCSE) for hierarchical Bayes (denoted by SDB, where $SDB = 100 \times |\frac{(posterior\ SD) - MCSE}{MCSE}|$). To answer research question 3 (regarding the estimation of posterior SD), the SDB will be compared between empirical Bayes and hierarchical Bayes with item covariates.

Hypotheses

Research question 1a: Accuracy of item parameter estimates. Because the use of group means results in shrinkage, which increases the accuracy of item parameter estimates (see p. 13), we expect both empirical Bayes and hierarchical Bayes to have lower RPB and lower RMSE than MMLE (which does not use group means at all, and therefore has no shrinkage). Because we expect MMLE to have high RMSE at small and medium sample sizes, we also expect empirical Bayes (which uses maximum likelihood estimates in Step 2) to have higher RPB and higher RMSE than hierarchical Bayes with item covariates (which does not use maximum likelihood estimates). Therefore, we expect the following relations regarding RPB and RMSE: empirical Bayes < MMLE, hierarchical Bayes with item covariates < MMLE, and hierarchical Bayes with item covariates < empirical Bayes.

Research question 1b: Acceptability of hierarchical Bayes. As discussed in the introduction, MMLE is expected to have difficulty with achieving convergence in smaller sample sizes, making estimation of item parameters impossible. In such conditions we expect a hierarchical Bayes method with item covariates to estimate item parameters and posterior SD with an acceptable degree of accuracy, having $RPB < 10\%$ and $SDB < 10\%$ for all item parameter types.

Research question 2: Added accuracy of item covariates. To examine the added value of item covariates in a hierarchical Bayes method for all conditions, we compare the accuracy of a hierarchical Bayes method both with and without item covariates. As discussed previously (see p. 13), because the use of multiple item covariates results in group shrinkage rather than total shrinkage, it is expected that hierarchical Bayes with item covariates

will have lower RPB than hierarchical Bayes without item covariates. Therefore, regarding RPB we expect hierarchical Bayes with item covariates < hierarchical Bayes without item covariates.[6]

Research question 3: Accuracy of posterior SD estimates. Because empirical Bayes ignores the uncertainty of item parameter estimates in Step 2, it is expected that empirical Bayes will result in large SDB. Alternatively, because hierarchical Bayes implements this uncertainty by using a one-step approach, it is expected that hierarchical Bayes will more accurately estimate its posterior SD than empirical Bayes, resulting in a smaller SDB. Therefore, for SDB we expect hierarchical Bayes with item covariates < empirical Bayes.

In addition to these hypotheses, certain patterns are expected across estimation methods regarding the three simulation factors:

Number of persons. It is expected that an increase in the number of persons will result in decreases in RPB and RMSE for these methods. As the number of persons approaches a medium sample size of 500, differences in the RMSE among these methods are expected to decrease as the accuracy of MMLE (which most prominently suffers in small sample sizes) increases. It is also expected than an increase in the number of persons will result in a decrease in SDB for both empirical Bayes and hierarchical Bayes.

Number of items. It is expected that an increase in the number of items will result in a decrease in the RPB as prior means are based on a larger number of items, therefore decreasing the shrinkage for individual items. Alternatively, it is expected that an increase in the number of items will result in higher RMSE, as there are more item parameters to estimate. For a fixed number of persons, it is expected that an increase in the number of items will result in higher SDB.

RSD. It is expected than an increase in RSD will result in a decrease in the RPB and an

[6]Although we expect lower RMSE when item covariates are used (in empirical Bayes and hierarchical Bayes methods) compared to when item covariates are not used (in MMLE), we have no such expectations for RMSE regarding the change in RMSE by using multiple item covariates (in hierarchical Bayes with item covariates) as opposed to using a single item covariates (in hierarchical Bayes without item covariates). As a result, we do not have any hypotheses regarding differences in RMSE between hierarchical Bayes with item covariates and hierarchical Bayes without item covariates.

increase in the RMSE, because less shrinkage is expected with a larger RSD. A larger RSD is not expected to affect SDB.

Results for Research Questions

For the majority of simulation conditions (23 out of 36), MMLE failed to converge for all 500 replications. Table D1 in Appendix D shows the percentage of 500 replications in which MMLE achieved convergence for each simulation condition. The most significant factor affecting convergence was RSD, with 83%, 25%, and 0% of conditions converging in all 500 replications with RSDs of 0.1, 0.3, and 0.5, respectively. Note that, because empirical Bayes estimates are calculated using maximum likelihood estimates, empirical Bayes estimates are unobtainable for those replications in which MMLE failed to converge. For the following analyses, only the 13 conditions in which MMLE had 100% convergence are considered for comparisons involving MMLE and/or empirical Bayes (i.e., research questions 1a and 3).

Research question 1a: Accuracy of item parameter estimates. Figure 3 and Table 6 present the RPB for each method (MMLE, empirical Bayes, and hierarchical Bayes with item covariates) in the 13 conditions that MMLE had 100% convergence. Each point in Figure 3 represents the maximum RPB for all item parameter types (α_i, β_{i1}, β_{i2}, β_{i3}, and β_{i4}), with each item parameter type averaged across replications.[7] Figure D1 and Tables D2-D4 in Appendix D present the full results for each item parameter type. Horizontal lines in Figure 3 indicate the cutoff for acceptable RPB of item parameter estimates (10%).

As shown in Figure 3, of the 13 conditions that MMLE had 100% convergence, MMLE had the lowest RPB of the three methods in 6 of those conditions (24 items and RSD=0.1 with 150, 200, 250, 300, and 500 persons [Figure 3, top-left], and 48 items and RSD = 0.3 with 500 persons [Figure 3, bottom-right]). Hierarchical Bayes had the lowest RPB in the other 7 conditions (24 items and RSD = 0.3 with 300 and 500 persons [Figure 3, top-right],

[7]We take this approach because we are uninterested in how accurately each method estimated individual item parameter types, but rather how accurately each method estimated *all* item parameter types. The maximum RPB for each condition indicates the range within which all item parameter types were estimated (for example, a value of 6% in Figure 3 indicates that all item parameter types for that condition were estimated by that method with $-6\% \leq RPB \leq 6\%$). This approach is used later on when presenting results for RMSE and SDB.

and 48 items and RSD = 0.1 with 150, 200, 250, 300, and 500 persons [Figure 3, bottom-left]). However, MMLE and hierarchical Bayes had highly comparable (within 2.02%) RPB in all 13 conditions. Empirical Bayes had the highest RPB of the three methods, only having lower RPB than MMLE in one condition (24 items and RSD = 0.3 with 500 persons), and in that condition the difference in RPB between the two methods was less than 0.23% (6.476% for MMLE, and 6.248% for empirical Bayes).

These results agree with our hypotheses regarding RPB that MMLE < empirical Bayes and hierarchical Bayes < empirical Bayes (as empirical Bayes consistently had the largest RPB of the three methods). However, results were highly similar between MMLE and hierarchical Bayes in these 13 conditions.

As shown in Table 6, empirical Bayes had unacceptably high RPB (RPB > 10%) in 11 conditions, only having acceptable RPB in 2 conditions (24 items and 0.3 RSD with 300 and 500 persons). MMLE and hierarchical Bayes had acceptably low RPB in all 13 conditions.

Figure 4 and Table 7 present the RMSE for each method in the 13 conditions that MMLE had 100% convergence. Each point in Figure 4 represents the maximum RMSE for all item parameter types, with each item parameter type averaged across replications. Figure D2 and Table D5-D7 in Appendix D present the full results for each item parameter type.

As shown in Figure 4, of the 13 conditions that MMLE had 100% convergence, empirical Bayes had the lowest RMSE in 3 conditions (24 items and RSD=0.3, with 300 and 500 persons [Figure 4, top-right], and 48 items and RSD=0.3, with 500 persons [Figure 4, bottom-right]). Hierarchical Bayes had the lowest RMSE in the other 10 conditions (24 items and RSD=0.1, with 150, 200, 250, 300, and 500 persons [Figure 4, top-left], and 48 items and RSD=0.1, with 150, 200, 250, 300, and 500 persons [Figure 4, bottom-left]). Empirical Bayes had the lowest RMSE of the three methods for the 3 conditions with RSD = 0.3 and had the highest RMSE of the three methods for the 10 conditions with RSD = 0.1. Hierarchical Bayes had lower RMSE than MMLE in all 13 conditions.

These results agree with our hypotheses regarding RMSE that hierarchical Bayes <

MMLE, and hierarchical Bayes < empirical Bayes (as hierarchical Bayes consistently had the smallest RMSE of the three methods). However, these results disagree with our hypothesis that empirical Bayes < MMLE, as MMLE had lower RMSE than empirical Bayes in 10 out of 13 conditions. Figure 4 illustrates that this unexpected result is likely because MMLE's accuracy decreased as the RSD increased, having an average RMSE of 0.357 when RSD=0.1 (Figure 4, left), and an average RMSE of 0.612 when RSD=0.3 (Figure 4, right), whereas empirical Bayes' performance improved (and surpassed MMLE) as RSD increases, having an average RMSE of 0.608 when RSD=0.1 (Figure 4, left), and an average RMSE of 0.550 when RSD=0.3 (Figure 4, right). It is possible that, had MMLE achieved 100% convergence in more conditions with RSD>0.1, we would see similar patterns of empirical Bayes outperforming MMLE in conditions with higher RSD.

To summarize, taking into account RPB and RMSE together, hierarchical Bayes with item covariates outperformed MMLE and empirical Bayes in the 13 conditions analyzed, having both RPB comparable to MMLE and generally lower RMSE than both MMLE and empirical Bayes.

Research question 1b: Acceptability of hierarchical Bayes. In the following we evaluate the acceptability of hierarchical Bayes with item covariates as an alternative to MMLE in the 23 conditions that MMLE failed to achieve 100% convergence. We examine the RPB and SDB of estimates obtained by hierarchical Bayes with item covariates in these conditions.[8]

Figure 5 and Table 8 present the RPB for hierarchical Bayes with covariates in the 23 conditions that MMLE failed to achieve convergence. Horizontal lines in Figure 5 indicate the cutoff for acceptable RPB of item parameter estimates (10%). As shown in Figure 5, hierarchical Bayes with covariates had acceptable RPB (< 10%) in 17 of the 23 conditions, having unacceptable RPB in the other 6 conditions (24 items and RSD = 0.3 with 100, 150,

[8]RMSE is not used as an evaluation measure for research question 1b because there is no single threshold for acceptable RMSE in these conditions, as RMSE is largely dependent on the level of RSD. However, the values for RMSE are still provided in Table 8.

and 200 persons [Figure 5, top-middle], 24 items and RSD = 0.5 with 100 persons [Figure 5, top-right], 48 items and RSD = 0.5 with 100 and 150 persons [Figure 5, bottom-right]). The primary factor affecting RPB in hierarchical Bayes with covariates was the number of persons, with hierarchical Bayes having acceptable RPB in all conditions with sample sizes ≥ 250 persons.

Figure 6 and Table 8 present the SDB for hierarchical Bayes with covariates in the 23 conditions that MMLE failed to achieve convergence. As shown in Figure 6 and Table 8, hierarchical Bayes with covariates had acceptable SDB ($< 10\%$) in 16 of the 23 conditions, having unacceptable SDB in the other 7 conditions (24 items and RSD = 0.3 with 100 persons [Figure 6, top-middle], 48 items and RSD = 0.1 with 100 persons [Figure 6, bottom-left], 48 items and RSD = 0.3 with 100, 150, 200, and 250 persons [Figure 6, bottom-middle], 48 items and RSD = 0.5 with 100 persons [Figure 6, bottom-right]). The primary factor affecting SDB in hierarchical Bayes with covariates was the number of items, having unacceptable SDB in one condition with 24 items and in six conditions with 48 items.

Taking both RPB and SDB into consideration, hierarchical Bayes with item covariates was an acceptable alternative to MMLE (having both $RPB < 10\%$ and $SDB < 10\%$) in 12 of the 23 conditions that MMLE failed to achieve convergence. Hierarchical Bayes showed to be an acceptable alternative to MMLE primarily in conditions with RSD = 0.5, with hierarchical Bayes being an acceptable alternative to MMLE in 9 of the 12 conditions with RSD = 0.5 (note that MMLE was unable to achieve 100% in *any* of the 12 conditions with RSD = 0.5).

Research question 2: Added accuracy of item covariates. Figure 7 and Table 9 present the RPB for hierarchical Bayes with covariates and hierarchical Bayes without covariates in all 36 conditions. Each point in Figure 7 represents the maximum RPB for all item parameter types, with each item parameter type averaged across replications. Horizontal lines in Figure 7 indicate the cutoff for acceptable RPB of item parameters (10%).

As shown in Figure 7, hierarchical Bayes without item covariates had lower RPB than

hierarchical Bayes with item covariates in only 9 of the 36 conditions (24 items and RSD = 0.3 with 500 persons, 24 items and RSD = 0.5 with 150, 200, 300, and 500 persons, 48 items and RSD = 0.3 with 100, 150, 200, and 500 persons), whereas hierarchical Bayes with item covariates had lower RPB in the other 27 conditions. Additionally, results were largely comparable between the two methods in these 9 conditions, with only one of these conditions having a difference in RPB larger than 2% (48 items and RSD = 0.3 with 100 persons, which had a difference in RPB of 2.66%). These results agree with our hypothesis regarding RPB that hierarchical Bayes with item covariates < hierarchical Bayes without item covariates, as hierarchical Bayes with item covariates had lower (or comparable) RPB to hierarchical Bayes without item covariates in all conditions.

Hierarchical Bayes with covariates had unacceptable RPB ($\geq$ 10%) in 7 of the 36 conditions (24 items and RSD = 0.3 with 100, 150, and 200 persons, 24 items and RSD = 0.5 with 100 persons, 48 items and RSD = 0.5 with 100, 500 persons), and hierarchical Bayes without covariates had unacceptable RPB in 16 of the 36 conditions (24 items and RSD = 0.3 with 100, 150, 200, 250, and 300 persons, 24 items and RSD = 0.5 with 100 persons, 48 items and RSD = 0.1 with 100, 150, 200, 250, 300, and 500 persons, 48 items and RSD = 0.5 with 100, 150, 200, and 300 persons). There were no conditions where hierarchical Bayes without covariates had acceptable RPB and hierarchical Bayes with item covariates had unacceptable RPB.

Figure 8 and Table 10 present the RMSE for hierarchical Bayes with covariates and hierarchical Bayes without covariates in all 36 conditions. Each point in Figure 8 represents the maximum RMSE for all item parameter types, with each item parameter type averaged across replications. As shown in Figure 8, hierarchical Bayes with item covariates had lower RMSE than hierarchical Bayes without item covariates in 13 of the 36 conditions. However, results were highly comparable between the two methods in 35 of the 36 conditions, with differences in RMSE < 0.06 for all but one condition. In one condition (48 items and RSD = 0.5 with 100 persons [Figure 8, bottom-right]) hierarchical Bayes without item covariates had

a significantly higher RMSE (1.364) than either method in any other condition. The most significant factor affecting RMSE for both methods was RSD, with both methods having larger RMSE as RSD increased.

In summary, hierarchical Bayes with covariates typically outperformed hierarchical Bayes without covariates, having lower (or comparable) RPB and lower (or comparable) RMSE in all 36 conditions.

Research question 3: Accuracy of posterior SD estimates. Figure 9 and Table 11 present the SDB for empirical Bayes and hierarchical Bayes with item covariates in the 13 conditions that MMLE had 100% convergence. Each point in Figure 9 represents the maximum SDB for all item parameter types, with each item parameter type averaged across replications. Figure D3 and Tables D8-D9 in Appendix D present the full results for each item parameter type.

As shown in Figure 9, of the 13 conditions that MMLE had 100% convergence, hierarchical Bayes with item covariates had lower SDB than empirical Bayes in 9 conditions (24 items and RSD = 0.1 with 150, 200, 250, 300, and 500 persons [Figure 9, top-left], 48 items and RSD = 0.1 with 150, 200, 250, and 300 persons [Figure 9, bottom-left]). Empirical Bayes had SDB comparable to hierarchical Bayes with item covariates in the remaining 4 conditions (24 items and RSD = 0.3 with 300 and 500 persons [Figure 9, top-right], 48 items and RSD = 0.1 with 500 persons [Figure 9, bottom-left], 48 items and RSD = 0.3 with 500 persons [Figure 9, bottom-right]).

These results agree with our hypothesis that, in general regarding SDB, hierarchical Bayes < empirical Bayes. Hierarchical Bayes outperformed empirical Bayes in conditions with RSD = 0.1, and results were highly comparable between the two methods for conditions with RSD = 0.3. A noticeable exception to these results is the condition having 48 items and RSD=0.1 with 500 persons (Figure 9, bottom-left), which had a sudden increase in SDB for hierarchical Bayes (increasing from 0.128 for 250 persons for 0.226 for 500 persons). This sudden increase resulted from a scaling artifact of SDB occurring when the MCSE in the denominator was

close to 0, despite posterior SD estimates and MCSE both decreasing with an increasing number of persons.[9]

As shown in Table 11, empirical Bayes had unacceptably high SDB (SDB > 10%) in the 10 conditions with RSD = 0.1, and acceptably low SDB in the 3 conditions with RSD = 0.3. Hierarchical Bayes had unacceptably high SDB in 8 conditions (24 items and RSD = 0.1 with 250, 300, and 500 persons, and 48 items and RSD = 0.1 with 150, 200, 250, 300, and 500 persons), and acceptably low SDB in the remaining 5 conditions (24 items and RSD = 0.1 with 150 and 200 persons, 24 items and RSD = 0.3 with 300 and 500 persons, and 48 items and RSD = 0.3 with 500 persons).

In summary, hierarchical Bayes with item covariates typically had lower SDB than empirical Bayes in the conditions that MMLE had 100% convergence. Hierarchical Bayes with item covariates had noticeably lower SDB in conditions with RSD = 0.1, and had highly similar results to empirical Bayes in conditions with RSD = 0.3.

Results Regarding Simulation Factors

Below, simulation results are interpreted in regards to the varying levels of the simulation factors.

Number of persons. Figures 3 and 7 show that RPB decreased with an increasing number of persons, especially for empirical Bayes with conditions having RSD = 0.1 (Figure 3, left), and for both hierarchical Bayes with and without item covariates in conditions with 24 items and RSD = 0.3, 24 items and RSD = 0.5, and 48 items and RSD = 0.5 (Figure 7, top-middle, top-right, and bottom-right).

Figure 4 shows that RMSE decreased with an increasing number of persons for empirical Bayes, especially for conditions having RSD = 0.1 (Figure 4, left). Figure 8 shows that the number of persons had little effect on the RMSE for both hierarchical Bayes with and without item covariates. An exception to this is the conditions with 48 items and RSD = 0.5 (Figure 8, bottom-right) for hierarchical Bayes without item covariates, where RMSE was

[9]Posterior SD estimates decreased from 0.084 to 0.065, and MCSE decreased from 0.075 to 0.053 as the number of persons increased from 250 to 500 persons.

especially high for hierarchical Bayes without item covariates with 100 persons.

Figure 9 shows that there is an interaction between the number of persons and the number of items regarding the SDB for both empirical Bayes and hierarchical Bayes with item covariates, with an increase in the number of persons resulting in an increase in SDB for conditions with 24 items and RSD = 0.1 (Figure 9, top-left), and a decrease in SDB (in general) for conditions with 48 items and RSD = 0.1 (Figure 9, bottom-left).

In summary, a larger number of persons is advisable when using empirical Bayes when items have small RSD (e.g., RSD = 0.1), and advisable for hierarchical Bayes (both with and without item covariates) when items have medium RSD (e.g., RSD = 0.3). However, a larger number of persons may result in large SDB if used in conjunction with a smaller number of items (e.g., 24).

Number of items. Figure 3 shows that an increase in the number of items resulted in a decrease in RPB for empirical Bayes. Figure 7 shows that an increase in the number of items resulted in an increase in RPB for hierarchical Bayes (both with and without item covariates) when RSD = 0.1 or RSD = 0.5, and a decrease in RPB when RSD = 0.3, implying that there was an interaction in regards to RPB between the number of items and the RSD.[10]

Figures 4 and 8 show that increasing the number of items resulted in a decrease in RMSE for empirical Bayes and a slight increase in RMSE for hierarchical Bayes with and without item covariates.

Figure 9 shows that an increase in the number of items resulted in higher SDB for both empirical Bayes and hierarchical Bayes with item covariates, increasing from an average SDB of 11.4% (in the 7 simulation conditions with 24 items) to 20.9% (in the 6 simulation conditions with 48 items) for empirical Bayes, and increasing from an average SDB of 7.8% (24 items) to 14.7% (48 items) for hierarchical Bayes with item covariates.

In summary, a larger number of items may be advisable for empirical Bayes to increase the

[10] A linear regression was imposed on the RPB for hierarchical Bayes with the number of persons, number of items, RSD, and their interactions as predictors. The statistically significant predictors of RPB were the number of items ($p < .005$), the RSD ($p < .001$), and the Items $\times$ RSD interaction ($p < .005$).

accuracy of item parameter estimates, at the cost of also increasing the SDB. Alternatively, a smaller number of items may be advisable to reduce RMSE and SDB in hierarchical Bayes. There were no clear patterns regarding the number of items and RPB for hierarchical Bayes.

RSD.[11] Figure 3 shows that an increase in RSD resulted in a decrease in RPB for empirical Bayes. Figure 7 shows that the relationship between RPB and RSD is highly dependent on the number of items for hierarchical Bayes with and without item covariates. For example, the highest RPB for 24 items occured when RSD = 0.3, whereas the lowest RPB for 48 items occured when RSD = 0.3.

Figure 4 shows that increasing RSD resulted in a decrease in RMSE for empirical Bayes, whereas Figure 8 shows that an increase in RSD resulted in an increase in RMSE for hierarchical Bayes with and without item covariates. Figure 9 shows that an increase in the RSD resulted in lower SDB for both empirical Bayes and for hierarchical Bayes with item covariates.

In summary, empirical Bayes tended to perform better (having lower RPB and SDB) in conditions with higher RSD. Hierarchical Bayes also had lower SDB in conditions with higher RSD, but had no clear pattern in RPB related to RSD.

Summary and Discussion

MMLE is the "gold standard" for estimating item parameters within an IRT framework. However, MMLE's accuracy, as well as its ability to achieve convergence, is limited in small sample sizes. Mislevy (1988) showed that auxiliary item information can be used to increase the accuracy of Rasch item location estimates with an empirical Bayes method. We presented hierarchical Bayes as an alternative to empirical Bayes both because RSD can be underestimated by empirical Bayes due to ignoring the uncertainty of item parameter estimates and because empirical Bayes is unable to obtain item parameter estimates when

[11]The following analyses regarding empirical Bayes are based on ten conditions for RSD = 0.1, three conditions for RSD = 0.3, and zero conditions for RSD = 0.5, as these are the 13 conditions in which MMLE achieved 100% convergence. The three conditions for RSD = 0.3 have sample sizes of 300, 500, and 500, making it difficult to distinguish between the effects of RSD and the number of persons on the evaluation measures between these conditions and the ten conditions for RSD = 0.1.

MMLE fails to achieve convergence. In this paper, we showed how item covariates can be used in empirical Bayes and hierarchical Bayes to obtain item parameter estimates of a GRM with higher accuracy and precision in small-to-medium sample sizes.

Method Selection Guideline

We provide a general guideline in Figure 10 based on simulation results regarding which method is recommended for different conditions. Table E1 in Appendix E presents a simplified version of the results previously discussed, regarding which method was considered the best (having the highest accuracy of item parameter estimates and posterior SD estimates) in each simulation condition examined. Because empirical Bayes failed to outperform hierarchical Bayes with item covariates in any of our simulation conditions (having either comparable or inferior accuracy of item parameter and posterior SD estimates in each condition), the only methods recommended in this section are MMLE and hierarchical Bayes with item covariates.

Step 1. The first step is to determine whether or not there are usable item covariates available and whether the test is unidimensional or multidimensional. If there are no item covariates available, then empirical Bayes and hierarchical Bayes with item covariates are not viable options. Because the empirical Bayes and hierarchical Bayes methods proposed in this study are based on the assumption that the test is unidimensional, we do not recommend these methods for a multidimensional test.

Step 2a. If there are no usable item covariates and/or the test is multidimensional, then if MMLE can achieve convergence we recommend using those estimates obtained using MMLE. However, if MMLE is unable to achieve convergence, we recommend either obtaining a larger sample size so that MMLE can achieve convergence or using alternative estimation methods.

Step 2b. If there are usable item covariates and the test is unidimensional, we make the following recommendations based on the number of items and the RSD of those items.

Step 3. To determine which method is recommended (given the availability of item

covariates and the number of items), estimates of the RSD are required. These estimates do not need to be highly accurate, but rather capable of allowing RSD to be categorizable as small (e.g., RSD=0.1), medium (e.g., RSD=0.3), or large (e.g., RSD=0.5). To obtain such estimates of the RSD, classical item discriminations and thresholds can be obtained to calculate linear regression RSD estimates. This method avoids the possibility of being unable to achieve convergence that MMLE is subject to, and can obtain results more efficiently than hierarchical Bayes. We provide an `R` function to calculate the linear regression RSD estimates based on classical item discriminations and thresholds and further discuss this function in Appendix B.

Step 3a. If there are a smaller number of items (e.g., 24), we make the following recommendations based on the RSD of items. If there is a small RSD (e.g., RSD=0.1, indicating items within groups are highly similar), we recommend using hierarchical Bayes with item covariates for sample sizes between 100 and 200, and MMLE for sample sizes > 200.[12] If there is a medium RSD (e.g., RSD=0.3, indicating that items within groups are similar yet distinctly different), we recommend hierarchical Bayes with item covariates for sample sizes ≥ 250. If there is a large RSD (e.g., RSD=0.5, indicating that items within groups are highly dissimilar), we recommend hierarchical Bayes with item covariates for sample sizes ≥ 150.

Step 3b. If there is a larger number of items (e.g., 48), we make the following recommendations based on the RSD of items. If there is a small RSD (e.g., RSD=0.1), we recommend MMLE for sample sizes ≥ 150. If there is a medium RSD (e.g., RSD=0.3), we recommend hierarchical Bayes with item covariates for sample sizes ≥ 300. If there is a large RSD (e.g., RSD = 0.5), we recommend hierarchical Bayes with item covariates for sample sizes ≥ 200.

Item Covariate Specification

As shown in this study, hierarchical Bayes with item covariates can be an acceptable alternative to MMLE under certain conditions. However, the effectiveness of hierarchical Bayes

[12] Although the maximum sample size used in this simulation study was 500 persons, we do not expect hierarchical Bayes or MMLE to have unacceptable results in sample sizes larger than 500.

is dependent on the correct specification of the item covariates structure. Both exploratory factor analysis and observation of the salient features of items are useful for assigning items to their correct groups and to assure a correct item covariate structure. Exploratory factor analysis can be used to identify how many dimensions (or domains within a single dimension) there are, and factor loadings can identify which items likely belong to each dimension/domain. The salient features of items (such as their similarities to other items with similar covariate structures) can be used to interpret these factors/dimensions in meaningful ways (e.g., identifying factors 1 and 2 as cognitive emotional fatigue and emotional listening fatigue) to make the classification of future items easier.

Mislevy (1988) illustrated how imposing a linear model on Rasch item location parameters based on item groupings can highlight misclassified items. Items with distinctly different properties from other items in their groups, such as an item with a significantly higher difficulty than any other item in its group, may indicate an incorrect item covariate structure. Looking at such items' salient features may show if (and how) it was misclassified, and what method of correcting the item covariate structure should be used. In Mislevy's (1988) empirical example, he shows three different methods that can be implemented to correct a misidentified item covariate structure: removing misfit items, creating a new item group, and changing the group status of certain items. Similar approaches can be applied to identifying and correcting mistakes in the item covariate structure of a GRM.

Study Limitations

This study had several methodological limitations that can be addressed in future research on these topics. First, a single item covariate structure (a mutually-exclusive binary Q-matrix with 6 item covariates and constant covariate effects across simulation conditions) was used in this simulation study to reflect the predominant covariate structure observed from an extensive literature review. In this study we also make the assumption that items are unidimensional, with item groups representing domains within a single underlying dimension. Future research using different item covariate structures, different effects of item covariates,

and generalizing these methods to allow multidimensionality may yield interesting results.

Second, in this study we assumed that the item covariate structure was correctly specified. The purpose of this study was to evaluate the added value of a correctly-identified item covariate structure through the use of empirical Bayes and hierarchical Bayes methods. The preliminary process of specifying the item covariate structure correctly is outside the scope of this study. Mislevy (1986) addressed how mispecifying the item covariate structure can result in "ensemble biases" affecting entire groups of items. Such biases can cause statistical properties (such as consistency) to no longer apply to item parameter estimates. Future research regarding the full repercussions of using an incorrect item covariate structure on empirical Bayes and hierarchical Bayes methods could be of interest.

Third, we used weakly informative priors for hierarchical Bayes in this study, as recommended by the statistical literature for small sample sizes. Using strongly informative priors can lead to bias in item parameter estimates, although the extent of this bias in small sample sizes may be of interest to future research.

Fourth, the levels selected for simulation factors (number of persons, number of items, and magnitude of RSD) reflect those we considered most relevant based on the literature. However, using additional levels of these simulation factors (e.g., 36 items, RSD=0.7) could show more clearly how evaluation criteria (RPB, RMSE, and SDB) change as a function of these simulation factors, such as comparing SDB for conditions with 24, 36, and 48 items.

Fifth, in this study we extended Mislevy's (1988) empirical Bayes method for a Rasch model to a GRM. An advantage of Mislevy's (1988) three-step approach is that the full item response data is not needed when MMLE is documented beforehand. However, the use of a three-step empirical Bayes method made it impossible to obtain results when MMLE was unable to converge. Because of this limitation, empirical Bayes and hierarchical Bayes could not be compared in the 23 simulation conditions of which MMLE was unable to achieve convergence, including all conditions with a sample size of 100 and/or RSD=0.5. Empirical Bayes, as it is most commonly used in the literature, is a one-step procedure similar in

implementation to hierarchical Bayes, but with different prior and posterior distribution specifications. Whereas a hierarchical Bayes method would allow hyperparameters to be estimated from hyper-prior distributions (e.g., the third line of Equation 26), an empirical Bayes method would treat these hyperparameters as fixed. Both a one-step empirical Bayes method and one-step hierarchical Bayes method could be implemented using MCMC (in software such as `rStan`), allowing results to be obtainable when MMLE is unable to achieve convergence. A one-step empirical Bayes method could be used in future research to allow empirical Bayes to be compared with hierarchical Bayes methods.

Conclusions

In this paper we have demonstrated the viability of empirical Bayes and hierarchical Bayes methods as alternatives to MMLE in small sample sizes. In addition, we have shown how to implement these methods using item covariates, and in what conditions these methods can result in acceptably accurate estimates of item parameters and RSD. Despite the aforementioned limitations of this study, we have demonstrated these methods and their implementation in conditions reflecting those most commonly found in the literature, and we have presented a framework that can be used in future research to expand upon these results under various other research conditions. In addition, we have provided the `R` functions written and utilized in this study to obtain empirical Bayes and hierarchical Bayes estimates for researchers to implement these proposed methods to their own research.

References

Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*(3), 251–269. https://doi.org/10.3102/10769986017003251

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17*(3), 201–210. https://doi.org/10.1177/014662169301700301

Baker, F. B. and Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bechger, T. M., Verstralen, H. H. F. M. & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika, 67*, 123–136. https://doi.org/10.1007/BF02294712

Bejar, I. I. (2012). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 50-66). New York, NY: Routledge. https://doi.org/10.4324/9780203803912

Beretvas, S. N., & Williams, N. J. (2004). The use of hierarchical generalized linear model for item dimensionality assessment. *Journal of Educational Measurement, 41*(4), 379-395. https://doi.org/10.1111/j.1745-3984.2004.tb01172.x

Betancourt, M. (2016). Identifying the optimal integration time in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.00225.*

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459. http://dx.doi.org/10.1007/BF02293801

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331-348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. http://doi:10.18637/jss.v048.i06

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement, 52*(2), 200-222. https://doi.org/10.1111/jedm.12072

Cho, S.-J., De Boeck, P., & Lee, W.-y. (2017). Evaluating testing, profile likelihood confidence interval estimation, and model comparisons for item covariate effects in linear logistic test models. *Applied Psychological Measurement, 41*(5), 353–371. http://doi.org/10.1177/0146621617692078

Cho, S.-J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2013). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika, 79*(1), 84–104. 10.1007/s11336-013-9360-2. http://doi.org/10.1007/s11336-013-9360-2

Choi, I. H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and psychological measurement, 75*(1), 78-101. https://doi.org/10.1177%2F0013164414522124

Curtis, S. (2010). BUGS code for item response theory. *Journal of Statistical Software, 36* (Code Snippet 1), 1 - 34. http://dx.doi.org/10.18637/jss.v036.c01

Davis, H., Schlundt, D., Camarata, S., Bess, F., & Hornsby, B. (2020, in review). Understanding listening-related fatigue: Perspectives of adults with hearing loss. *International Journal of Audiology.*

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533. https://doi.org/10.1007/s11336-008-9092-x

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195*(2), 216-222. https://doi.org/10.1016/0370-2693(87)91197-X

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75,* 474-497. https://doi.org/10.1007/s11336-010-9161-9

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 300–396. https://psycnet.apa.org/doi/10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433. https://doi.org/10.1007/BF02294564

Embretson, S. E. (2015). The multicomponent latent trait model for diagnosis: Applications to heterogeneous test domains. *Applied Psychological Measurement, 39*(1), 16-30. https://doi.org/10.1177%2F0146621614552014

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6*(4),

397-416. https://doi.org/10.1177%2F014662168200600403

Fischer, H. F., & Rose, M. (2019). Scoring depression on a common metric: A comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivariate Behavioral Research, 54*(1), 85-99. https://doi.org/10.1080/00273171.2018.1491381

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275–299. http://doi.org/10.1037/a0015825

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-0742-4

Freund, P. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement, 32*(3), 195-210. https://doi.org/10.1177%2F0146621607306972

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515-534. http://doi.org/10.1214/06-BA117A

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511. http://doi.org/10.1214/ss/1177011136

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (163-174). Boca Raton, FL: Taylor & Francis. http://doi.org/10.1201/b10905

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: Chapman & Hall.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4), 351-373. https://doi.org/10.1111/j.1745-3984.2005.00020.x

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*(4), 665-686. https://doi.org/10.1177%2F0013164411430707

Hoffman, L., Yang, X., Bovaird, J. A., & Embretson, S. E. (2006). Measuring attentional ability in older adults: Development and psychometric evaluation of DriverScan. *Educational and Psychological Measurement, 66*(6), 984-1000. https://doi.org/10.1177%2F0013164406288170

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593-1623. https://arxiv.org/abs/1111.4246

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746. https://doi.org/10.1177%2F0013164410390032

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*(4), 369-380. https://doi.org/10.1177%2F014662168601000405

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. http://doi.org/10.1080/10705519909540118

Ip, E. H., Magee, M. F., Youssef, G. A., & Chen, S. H. (2019). Gleaning information for cognitive operations from don't know responses in cognitive

and noncognitive assessments. *Multivariate Behavioral Research, 54*(2), 159-172. https://doi.org/10.1080/00273171.2018.1503075

Ip, E. H., Smits, D. J. M., & De Boeck, P. (2009). Locally dependent linear logistic test model with person covariates. *Applied Psychological Measurement, 33*(7), 555-569. https://doi.org/10.1177/0146621608326424

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499-518. https://doi.org/10.1177%2F0146621608327800

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36,* 399-419. https://doi.org/10.1177%2F0146621612446170

Kim, J. (2018). *Extensions and applications of item explanatory models to polytomous data in item response theory* (Doctoral dissertation, UC Berkeley). Retrieved from https://escholarship.org/uc/item/4pf2p2fj

Kubinger, K. D. (2008). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232–244. https://doi.org/10.1177/0013164408322021

Lautenschlager, G. J., Meade, A. W., & Kim, S.-H. (2006, April). *Cautions regarding sample characteristics when using the graded response model.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Medina-Díaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement, 17*(2), 117-130. https://doi.org/10.1177%2F014662169301700202

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195. https://doi.org/10.1007/BF02293979

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12*(3), 281-296. https://doi.org/10.1177%2F014662168801200306

Mitchell, K. J. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the armed services vocational aptitude battery.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics, 111*(1), 194-203. https://arxiv.org/abs/hep-lat/9208011

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.) *Handbook of Markov chain Monte Carlo* (pp. 113-162). Boca Raton, FL: Chapman & Hall. https://doi.org/10.1080/09332480.2012.668472

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178. https://doi.org/10.2307/1165199

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly, 51*(2), 123-134.

Polson, N. G., & Scott, S. L. (2011). Data augmentation for support vector machines. *Bayesian Analysis, 6*(1), 1-23. https://doi.org/10.1214/11-BA601

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Rakkapao, S., Prasitpong, S., & Arayathanitkul, K. (2016). Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique. *Physical Review Physics Education Research, 12*(2), 020135. https://doi.org/10.1103/PhysRevPhysEducRes.12.020135

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x

Rencher, A. C. (2000). *Linear models in statistics.* New York, NY: Wiley.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*(1), 42-56. https://doi.org/10.1177%2F0146621602026001003

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(S1), 1–97. http://doi.org/10.1007/BF03372160

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*(3), 255–272. https://doi.org/10.1111/j.1745-3984.1990.tb00747.x

Shermis, M. D., & Chang, S.-H. (1997). The use of item response theory (IRT) to investigate the hierarchical nature of a college mathematics curriculum. *Educational and Psychological Measurement, 57*(3), 450-458. https://doi.org/10.1177%2F0013164497057003006

Stan Development Team. (2018). RStan: the R interface to Stan. R package version 2.17.3. Retrieved from http://mc-stan.org

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5*(3), 383-397. https://doi.org/10.1177%2F014662168100500312

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California, Los Angeles). Retrieved from https://www.statmodel.com/download/Yudissertation.pdf

Table 1: *Fit Indices from Exploratory Factor Analyses*

Fix Indices	1-Factor	2-Factor	3-Factor	4-Factor
SRMR	0.038	0.028	0.022	0.018
RMSEA	0.082[0.079,0.085]*	0.072[0.069,0.075]*	0.062[0.059,0.064]*	0.054[0.051,0.057]*
CFI	0.986	0.990	0.993	0.995
TLI	0.985	0.989	0.992	0.994

Note. * 90% confidence interval

Table 2: *Regression Coefficients for Empirical Bayes vs. Hierarchical Bayes, J=150*

	Empirical Bayes		Hierarchical Bayes				
	EST	SE	Mean	Median	SD	0.025*	0.975*
Discrimination							
$\gamma_{\alpha 0}$	2.095	0.251	2.143	2.135	0.281	1.620	2.719
$\gamma_{\alpha 1}$	0.673	0.365	0.509	0.509	0.361	-0.229	1.208
$\gamma_{\alpha 2}$	0.979	0.355	0.999	0.993	0.368	0.285	1.736
$\gamma_{\alpha 3}$	0.458	0.355	0.319	0.313	0.355	-0.380	1.019
ϕ_{α}	0.794	0.147	0.741	0.732	0.112	0.546	0.986
Threshold 1							
$\gamma_{\beta 01}$	-4.149	0.593	-4.001	-3.989	0.558	-5.095	-2.912
$\gamma_{\beta 11}$	-2.116	0.861	-2.168	-2.173	0.782	-3.742	-0.632
$\gamma_{\beta 21}$	-1.356	0.838	-1.303	-1.317	0.767	-2.777	0.200
$\gamma_{\beta 31}$	0.142	0.838	0.500	0.510	0.767	-1.023	2.046
$\phi_{\beta 1}$	1.874	0.820	1.619	1.602	0.236	1.216	2.135
Threshold 2							
$\gamma_{\beta 02}$	-2.179	0.423	-1.997	-2.000	0.423	-2.795	-1.175
$\gamma_{\beta 12}$	-1.874	0.615	-1.896	-1.889	0.571	-3.050	-0.775
$\gamma_{\beta 22}$	-1.041	0.598	-0.973	-0.966	0.552	-2.090	0.137
$\gamma_{\beta 32}$	0.548	0.598	0.714	0.727	0.551	-0.395	1.767
$\phi_{\beta 2}$	1.338	0.418	1.202	1.187	0.162	0.928	1.558
Threshold 3							
$\gamma_{\beta 03}$	-0.249	0.353	-0.071	-0.073	0.403	-0.865	0.717
$\gamma_{\beta 13}$	-1.238	0.513	-1.409	-1.410	0.527	-2.426	-0.379
$\gamma_{\beta 23}$	-0.522	0.499	-0.461	-0.466	0.527	-1.486	0.573
$\gamma_{\beta 33}$	0.722	0.499	0.772	0.763	0.523	-0.253	1.775
$\phi_{\beta 3}$	1.116	0.290	1.110	1.098	0.144	0.867	1.435
Threshold 4							
$\gamma_{\beta 04}$	1.653	0.314	1.811	1.815	0.382	1.070	2.555
$\gamma_{\beta 14}$	-0.381	0.456	-0.627	-0.631	0.492	-1.598	0.345
$\gamma_{\beta 24}$	0.081	0.444	0.158	0.162	0.497	-0.825	1.109
$\gamma_{\beta 34}$	1.009	0.444	0.970	0.967	0.502	-0.022	1.932
$\phi_{\beta 4}$	0.993	0.230	1.067	1.051	0.142	0.835	1.382

Note. * Percentiles of posterior distribution

Table 3: *Comparison of Empirical and Hierarchical Bayesian Estimates for α_i with J=150*

		Empirical							Hierarchical	
		Step 1		Step 2		Step 3				
Item	Domain	$\hat{\alpha}_i$	$\hat{\tau}_{\alpha i}$	$\bar{\alpha}_i$	ϕ_α	$\tilde{\alpha}_i$	$\tilde{\sigma}_{\alpha i}$	Shrinkage	Posterior Median	SD
1	C	-5.361	0.802	NA	NA	NA	NA	NA	-7.180	1.263
2	C	-8.430	1.250	-6.265	1.874	-7.763	1.040	0.308	-7.628	0.925
3	C	-8.382	1.068	-6.265	1.874	-7.863	0.928	0.245	-7.267	0.760
4	C	-6.893	1.121	-6.265	1.874	-6.728	0.962	0.263	-6.701	0.936
5	C	-7.202	0.921	-6.265	1.874	-7.020	0.826	0.194	-6.705	0.726
6	C	-5.978	0.723	-6.265	1.874	-6.015	0.674	0.129	-6.056	0.660
7	C	-4.747	0.594	-6.265	1.874	-4.886	0.567	0.091	-4.573	0.532
8	C	-5.924	0.758	-6.265	1.874	-5.972	0.702	0.141	-6.094	0.702
9	C	-4.927	0.596	-6.265	1.874	-5.050	0.568	0.092	-4.984	0.531
10	C	-3.901	0.458	-6.265	1.874	-4.034	0.445	0.056	-4.156	0.478
11	E	-4.324	0.556	-5.505	1.874	-4.419	0.533	0.081	-4.600	0.575
12	E	-5.863	0.704	-5.505	1.874	-5.819	0.659	0.124	-5.657	0.622
13	E	-4.778	0.565	-5.505	1.874	-4.838	0.541	0.083	-4.920	0.554
14	E	-5.523	0.674	-5.505	1.874	-5.521	0.635	0.115	-5.271	0.575
15	E	-6.735	0.820	-5.505	1.874	-6.537	0.751	0.161	-6.273	0.654
16	E	-3.166	0.456	-5.505	1.874	-3.297	0.443	0.056	-3.139	0.424
17	E	-6.319	0.757	-5.505	1.874	-6.205	0.702	0.140	-5.966	0.623
18	E	-7.211	0.880	-5.505	1.874	-6.903	0.797	0.181	-6.536	0.695
19	E	-5.848	0.755	-5.505	1.874	-5.800	0.700	0.140	-5.380	0.599
20	E	-5.284	0.632	-5.505	1.874	-5.307	0.599	0.102	-5.156	0.553
21	P	-8.197	1.044	-4.007	1.874	-7.205	0.912	0.237	-6.644	0.738
22	P	-7.240	0.937	-4.007	1.874	-6.594	0.838	0.200	-6.120	0.688
23	P	-2.773	0.392	-4.007	1.874	-2.825	0.383	0.042	-2.695	0.369
24	P	-7.667	0.966	-4.007	1.874	-6.898	0.859	0.210	-6.150	0.657
25	P	-5.524	0.692	-4.007	1.874	-5.342	0.649	0.120	-4.766	0.539
26	P	-2.033	0.319	-4.007	1.874	-2.089	0.314	0.028	-1.962	0.297
27	P	-1.330	0.228	-4.007	1.874	-1.369	0.226	0.015	-1.357	0.230
28	P	-2.736	0.389	-4.007	1.874	-2.789	0.381	0.041	-2.640	0.358
29	P	-1.115	0.237	-4.007	1.874	-1.160	0.235	0.016	-1.124	0.230
30	P	-1.453	0.287	-4.007	1.874	-1.511	0.284	0.023	-1.413	0.270
31	S	-6.171	0.841	-4.149	1.874	-5.832	0.767	0.168	-5.709	0.704
32	S	-3.524	0.413	-4.149	1.874	-3.553	0.404	0.046	-3.535	0.391
33	S	-6.904	0.849	-4.149	1.874	-6.435	0.773	0.170	-6.057	0.662
34	S	-4.947	0.595	-4.149	1.874	-4.874	0.567	0.091	-4.620	0.527
35	S	-3.422	0.406	-4.149	1.874	-3.454	0.396	0.045	-3.430	0.383
36	S	-3.266	0.401	-4.149	1.874	-3.304	0.392	0.044	-3.242	0.387
37	S	-3.367	0.448	-4.149	1.874	-3.409	0.436	0.054	-3.173	0.401
38	S	-3.218	0.389	-4.149	1.874	-3.257	0.381	0.041	-3.239	0.376
39	S	-3.026	0.390	-4.149	1.874	-3.073	0.382	0.042	-2.964	0.367
40	S	-3.647	0.430	-4.149	1.874	-3.672	0.419	0.050	-3.718	0.411

Note. Maximum likelihood estimation was unable to estimate the fourth threshold of item 1 ($\beta_{1,4}$) because no responses were obtained in the fifth category of item 1. Because of this, item 1 was omitted during Step 2 when obtaining regression estimates, resulting in the missing values for item 1 (noted as NA). Because regression estimates were required for the values calculated in Step 3, these values are also missing for item 1. This is one instance where hierarchical Bayes estimation has an advantage over empirical Bayes estimation. Because the hyper-prior distribution for $\beta_{1,4}$ in hierarchical Bayesian estimation provides information outside of the data, $\beta_{1,4}$ is still capable of being estimated.

Table 4: *Comparison of Empirical and Hierarchical Bayesian Estimates for $\beta_{i,4}$ with J=150*

		Empirical							Hierarchical	
		Step 1		Step 2		Step 3				
Item	Domain	$\widehat{\beta}_{i,4}$	$\widehat{\tau}_{\beta i,4}$	$\bar{\beta}_{i,4}$	$\bar{\phi}_{\beta,4}$	$\tilde{\beta}_{i,4}$	$\tilde{\sigma}_{\beta i,4}$	Shrinkage	Posterior Median	SD
1	C	NA	NA	NA	NA	NA	NA	NA	-1.184	0.247
2	C	0.712	0.352	1.272	0.993	0.774	0.332	0.111	0.927	0.312
3	C	2.128	0.499	1.272	0.993	1.956	0.446	0.202	2.150	0.400
4	C	-0.306	0.285	1.272	0.993	-0.186	0.274	0.076	-0.067	0.257
5	C	0.611	0.368	1.272	0.993	0.691	0.345	0.121	0.829	0.320
6	C	1.347	0.339	1.272	0.993	1.339	0.320	0.104	1.548	0.323
7	C	3.144	0.483	1.272	0.993	2.785	0.435	0.191	3.154	0.421
8	C	1.366	0.312	1.272	0.993	1.358	0.298	0.090	1.555	0.301
9	C	1.079	0.341	1.272	0.993	1.100	0.323	0.105	1.306	0.326
10	C	1.364	0.280	1.272	0.993	1.357	0.270	0.074	1.522	0.283
11	E	0.417	0.215	1.735	0.993	0.476	0.210	0.045	0.586	0.221
12	E	2.078	0.392	1.735	0.993	2.032	0.365	0.135	2.287	0.361
13	E	1.659	0.315	1.735	0.993	1.666	0.300	0.091	1.913	0.316
14	E	1.396	0.409	1.735	0.993	1.445	0.378	0.145	1.682	0.371
15	E	2.127	0.446	1.735	0.993	2.061	0.407	0.168	2.307	0.397
16	E	2.875	0.429	1.735	0.993	2.696	0.394	0.157	3.058	0.400
17	E	1.822	0.421	1.735	0.993	1.809	0.387	0.152	2.036	0.382
18	E	2.297	0.478	1.735	0.993	2.191	0.431	0.188	2.432	0.410
19	E	1.093	0.461	1.735	0.993	1.207	0.418	0.177	1.405	0.401
20	E	1.581	0.391	1.735	0.993	1.602	0.364	0.134	1.862	0.370
21	P	2.213	0.451	2.662	0.993	2.290	0.411	0.171	2.268	0.378
22	P	-0.226	0.358	2.662	0.993	0.107	0.337	0.115	0.210	0.284
23	P	2.399	0.365	2.662	0.993	2.430	0.342	0.119	2.647	0.347
24	P	3.550	0.562	2.662	0.993	3.335	0.489	0.242	3.315	0.434
25	P	3.128	0.519	2.662	0.993	3.028	0.460	0.214	3.138	0.433
26	P	3.986	0.467	2.662	0.993	3.747	0.422	0.181	3.998	0.418
27	P	2.201	0.279	2.662	0.993	2.235	0.269	0.073	2.375	0.283
28	P	2.903	0.395	2.662	0.993	2.870	0.367	0.137	3.105	0.378
29	P	2.873	0.345	2.662	0.993	2.850	0.326	0.108	2.996	0.328
30	P	3.590	0.427	2.662	0.993	3.446	0.392	0.156	3.654	0.381
31	S	0.700	0.281	1.653	0.993	0.771	0.270	0.074	0.909	0.265
32	S	1.423	0.280	1.653	0.993	1.440	0.269	0.074	1.628	0.271
33	S	1.047	0.369	1.653	0.993	1.121	0.346	0.122	1.252	0.313
34	S	-0.269	0.282	1.653	0.993	-0.126	0.271	0.075	0.004	0.253
35	S	2.024	0.300	1.653	0.993	1.993	0.287	0.084	2.181	0.295
36	S	2.067	0.319	1.653	0.993	2.028	0.304	0.094	2.251	0.304
37	S	3.449	0.446	1.653	0.993	3.147	0.407	0.168	3.411	0.405
38	S	2.083	0.301	1.653	0.993	2.047	0.288	0.084	2.234	0.298
39	S	2.801	0.373	1.653	0.993	2.659	0.349	0.123	2.859	0.342
40	S	1.206	0.257	1.653	0.993	1.234	0.248	0.063	1.396	0.251

Note. Maximum likelihood estimation was unable to estimate the fourth threshold of item 1 ($\beta_{1,4}$) because no responses were obtained in the fifth category of item 1. Because of this, item 1 was omitted during Step 2 when obtaining regression estimates, resulting in the missing values for item 1 (noted as NA).

Table 5: *LLTM Literature Review*

Reference	Number of Items	Number of Item Covariates (per factor)	Item Covariate Structure
Baker (1993)	21	8	Non-mutually exclusive binary Q-matrix
Bechger, Verstralen, & Verhelst (2002)	5	2	Non-mutually exclusive non-binary Q-matrix
Beretvas & Williams (2004)	17	2	Mutually exclusive binary Q-matrix
Bolt, Cohen, & Wollack (2002)	26	2	Mutually exclusive binary Q-matrix
Chalmers (2015)	15	3	Mutually exclusive binary Q-matrix
Choi & Wilson (2015)	24	2x2x3=12**	Q-matrix by factor
De Boeck (2008)	24	2x2x3=12**	Q-matrix by factor
Embretson (2015)	70	4	Non-mutually exclusive binary Q-matrix
Fischer (1973)	29	8	Non-mutually exclusive binary Q-matrix
Freund, Hofer, & Holling (2008)	25	5	Non-mutually exclusive binary Q-matrix
Gorin (2005)	29	5	Mutually exclusive binary Q-matrix
Hartig, Frey, Nold, & Klieme (2012)	46	2x2=4**	Q-matrix by factor
Hoffman, Yang, Bovaird, & Embretson (2006)	64	3	Non-mutually exclusive non-binary Q-matrix
Hohensinn & Kubinger (2011)	18	3	Mutually exclusive binary Q-matrix
Hornke & Habon (1986)	24	8x3=24**	Q-matrix by factor
Ip, Magee, Youssef, & Chen (2019)	31	6	Non-mutually exclusive binary Q-matrix
Ip, Smits, & De Boeck (2009)	8	2x2=4**	Q-matrix by factor
Kim (2018)	13	3x3x3=27**	Q-matrix by factor
Kubinger (2008)	29	8	Non-mutually exclusive binary Q-matrix
Medina-Diaz (1993)	29	8	Non-mutually exclusive binary Q-matrix
Mislevy (1988)	20	6	Mutually exclusive binary Q-matrix
Mitchell (1983)	334	10	Mutually exclusive binary Q-matrix
Poinstingl (2009)	25	8***	Q-matrix by factor
Rakkapao, Prasitpong, & Arayathanitkul (2016)	20	10	Mutually exclusive binary Q-matrix
Rost & Cartensen (2002)	77	11x7=77**	Q-matrix by factor
Sheehan & Mislevy (1990)	93	3	Mutually exclusive binary Q-matrix
Shermis & Chang (1997)	45/90/90*	4	Mutually exclusive binary Q-matrix
Whitely & Schneider (1981)	30	8	Non-mutually exclusive non-binary Q-matrix

Note. * Three forms of the same test were used in this study. Form A had 45 items, and Forms B/C each had 90 items.
** The number of item groups is equal to the product of the number of levels per factor. For example, three factors with two levels for each factor results in $2 \times 2 \times 2 = 8$ item groups.
*** The number of factors in this study was 8. Three factors had mutually exclusive binary Q-matrices with 4, 3, and 5 levels. The remaining five factors counted the occurrences of different item attributes.

Table 6: *Research Question 1a: RPB Comparison for MMLE, Empirical Bayes, and Hierarchical Bayes*

# Items	RSD	# Persons	MMLE	EB	HB
24	0.1	150	2.739	63.628	4.119
24	0.1	200	1.872	59.218	3.393
24	0.1	250	1.417	58.262	2.392
24	0.1	300	1.222	54.943	2.007
24	0.1	500	0.947	46.576	1.281
24	0.3	300	7.129	9.344	6.570
24	0.3	500	6.476	6.248	4.456
48	0.1	150	6.117	50.321	5.929
48	0.1	200	5.751	47.551	5.551
48	0.1	250	5.443	44.373	4.987
48	0.1	300	5.045	41.550	4.850
48	0.1	500	4.753	33.500	4.679
48	0.3	500	2.609	10.690	3.058

Table 7: *Research Question 1a: RMSE Comparison for MMLE, Empirical Bayes, and Hierarchical Bayes*

# Items	RSD	# Persons	MMLE	EB	HB
24	0.1	150	0.392	0.814	0.278
24	0.1	200	0.357	0.732	0.263
24	0.1	250	0.334	0.701	0.253
24	0.1	300	0.320	0.662	0.247
24	0.1	500	0.290	0.518	0.238
24	0.3	300	0.616	0.534	0.547
24	0.3	500	0.596	0.543	0.553
48	0.1	150	0.428	0.677	0.295
48	0.1	200	0.393	0.624	0.285
48	0.1	250	0.370	0.578	0.276
48	0.1	300	0.356	0.536	0.271
48	0.1	500	0.326	0.420	0.268
48	0.3	500	0.623	0.573	0.576

Table 8: *Research Question 1b: Acceptability of Hierarchical Bayes with Covariates as an Alternative to MMLE*

# Items	RSD	# Persons	RPB	RMSE	SDB	Acceptable Substitute?
24	0.1	100	6.139	0.303	5.876	Yes
24	0.3	100	25.967	0.550	11.054	No
24	0.3	150	18.424	0.551	8.654	No
24	0.3	200	15.172	0.548	8.545	No
24	0.3	250	9.757	0.547	9.774	Yes
24	0.5	100	15.113	0.948	4.649	No
24	0.5	150	9.757	0.937	5.461	Yes
24	0.5	200	7.090	0.934	5.699	Yes
24	0.5	250	4.982	0.936	4.318	Yes
24	0.5	300	4.547	0.936	3.357	Yes
24	0.5	500	2.824	0.933	1.302	Yes
48	0.1	100	5.876	0.304	14.173	No
48	0.3	100	9.077	0.559	21.465	No
48	0.3	150	6.193	0.563	12.525	No
48	0.3	200	5.153	0.562	16.209	No
48	0.3	250	3.835	0.566	12.966	No
48	0.3	300	3.231	0.568	7.207	Yes
48	0.5	100	15.632	0.884	12.105	No
48	0.5	150	12.980	0.916	4.829	No
48	0.5	200	9.595	0.925	6.034	Yes
48	0.5	250	8.925	0.933	4.363	Yes
48	0.5	300	9.575	0.938	4.229	Yes
48	0.5	500	7.524	0.952	2.658	Yes

Table 9: *Research Question 2: RPB Comparison for Hierarchical Bayes with Covariates (HB with) and Hierarchical Bayes without Covariates (HB without)*

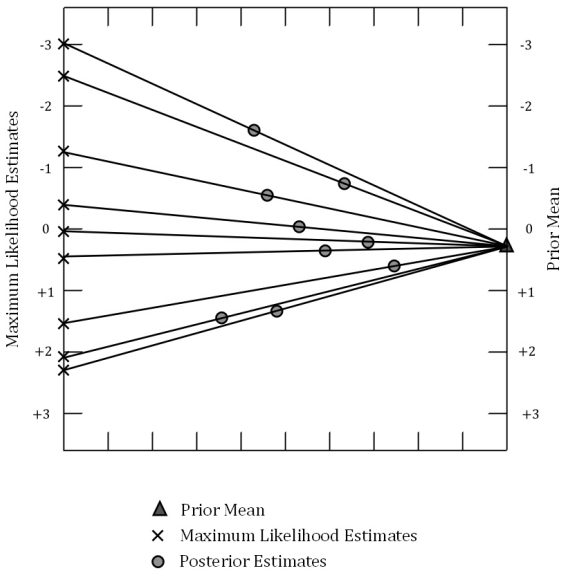
# Items	RSD	# Persons	HB with	HB without
24	0.1	100	6.139	9.689
24	0.1	150	4.119	8.041
24	0.1	200	3.393	8.046
24	0.1	250	2.392	6.359
24	0.1	300	2.007	5.878
24	0.1	500	1.281	4.344
24	0.3	100	25.967	45.848
24	0.3	150	18.424	30.405
24	0.3	200	15.172	23.153
24	0.3	250	9.757	15.714
24	0.3	300	6.570	11.078
24	0.3	500	4.456	4.420
24	0.5	100	15.113	16.341
24	0.5	150	9.757	8.156
24	0.5	200	7.090	5.659
24	0.5	250	4.982	6.084
24	0.5	300	4.547	3.681
24	0.5	500	2.824	2.373
48	0.1	100	5.876	17.162
48	0.1	150	5.929	16.140
48	0.1	200	5.551	15.163
48	0.1	250	4.987	14.903
48	0.1	300	4.850	13.673
48	0.1	500	4.679	11.078
48	0.3	100	9.077	6.418
48	0.3	150	6.193	5.383
48	0.3	200	5.153	4.309
48	0.3	250	3.835	4.139
48	0.3	300	3.231	3.838
48	0.3	500	3.058	3.051
48	0.5	100	15.632	30.358
48	0.5	150	12.980	23.525
48	0.5	200	9.595	11.942
48	0.5	250	8.925	9.736
48	0.5	300	9.575	10.117
48	0.5	500	7.539	7.870

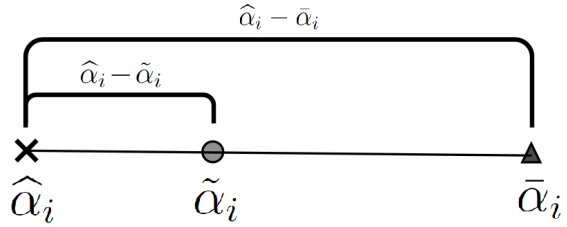
Table 10: *Research Question 2: RMSE Comparison for Hierarchical Bayes with Covariates (HB with) and Hierarchical Bayes without Covariates (HB without)*

# Items	RSD	# Persons	HB with	HB without
24	0.1	100	0.303	0.329
24	0.1	150	0.278	0.312
24	0.1	200	0.263	0.303
24	0.1	250	0.253	0.291
24	0.1	300	0.247	0.284
24	0.1	500	0.238	0.266
24	0.3	100	0.550	0.511
24	0.3	150	0.551	0.530
24	0.3	200	0.548	0.537
24	0.3	250	0.547	0.539
24	0.3	300	0.547	0.541
24	0.3	500	0.553	0.549
24	0.5	100	0.948	0.892
24	0.5	150	0.937	0.904
24	0.5	200	0.934	0.914
24	0.5	250	0.936	0.917
24	0.5	300	0.936	0.920
24	0.5	500	0.933	0.924
48	0.1	100	0.304	0.329
48	0.1	150	0.295	0.313
48	0.1	200	0.285	0.305
48	0.1	250	0.276	0.301
48	0.1	300	0.271	0.295
48	0.1	500	0.268	0.282
48	0.3	100	0.559	0.521
48	0.3	150	0.563	0.536
48	0.3	200	0.562	0.544
48	0.3	250	0.566	0.553
48	0.3	300	0.568	0.557
48	0.3	500	0.576	0.571
48	0.5	100	0.884	1.364
48	0.5	150	0.916	0.909
48	0.5	200	0.925	0.922
48	0.5	250	0.933	0.928
48	0.5	300	0.938	0.934
48	0.5	500	0.953	0.950

Table 11: *Research Question 3: SDB Comparison for Empirical Bayes and Hierarchical Bayes*

# Items	RSD	# Persons	EB	HB
24	0.1	150	10.202	4.316
24	0.1	200	12.988	6.980
24	0.1	250	17.640	10.101
24	0.1	300	14.543	11.119
24	0.1	500	14.387	10.508
24	0.3	300	6.029	7.071
24	0.3	500	4.257	4.735
48	0.1	150	29.738	15.962
48	0.1	200	26.812	13.817
48	0.1	250	23.463	13.295
48	0.1	300	21.265	14.982
48	0.1	500	15.967	22.554
48	0.3	500	8.336	7.656





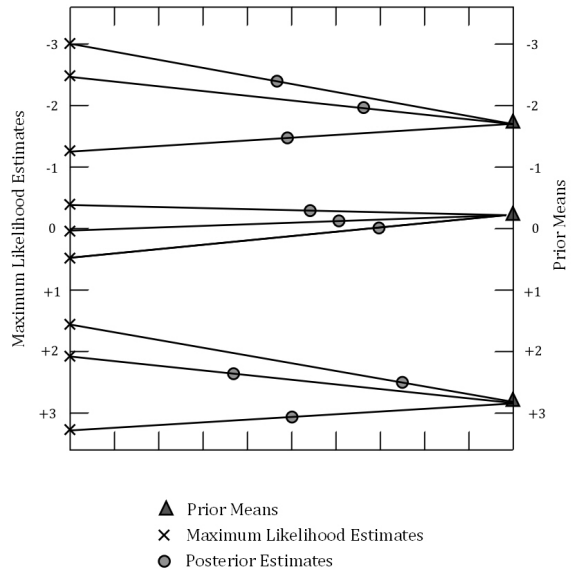
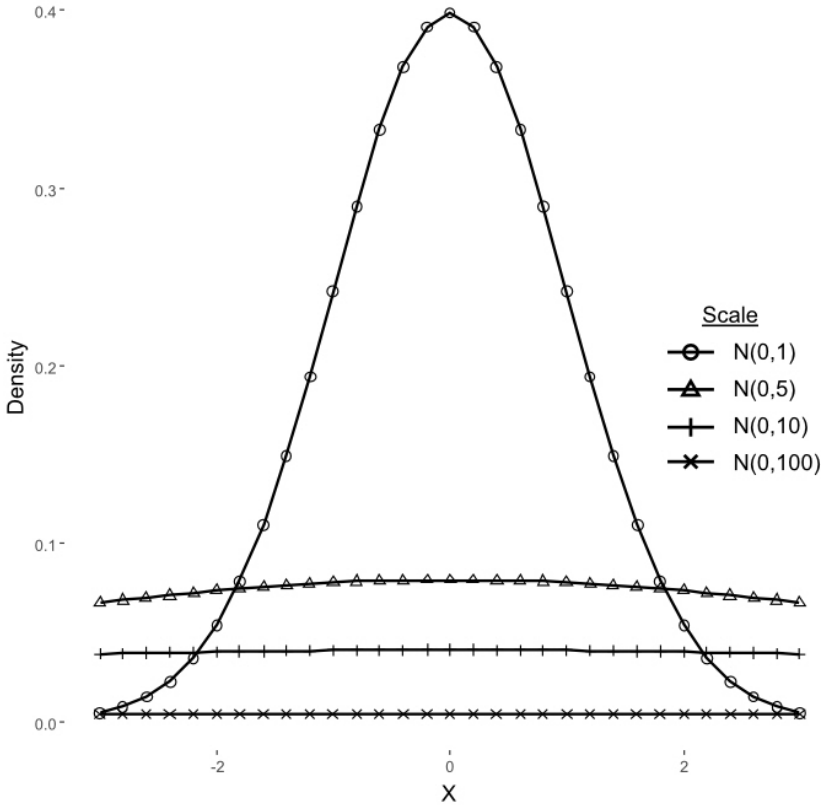


Figure 1: Illustration of shrinkage



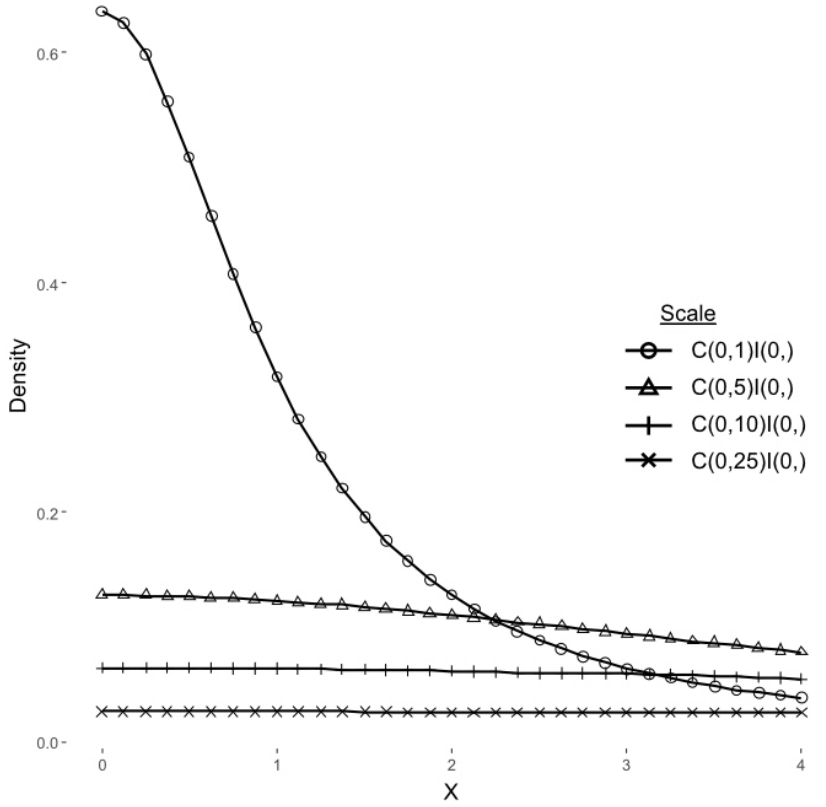


Figure 2: Probability density functions of different normal (top) and half-Cauchy (bottom) distribution scales

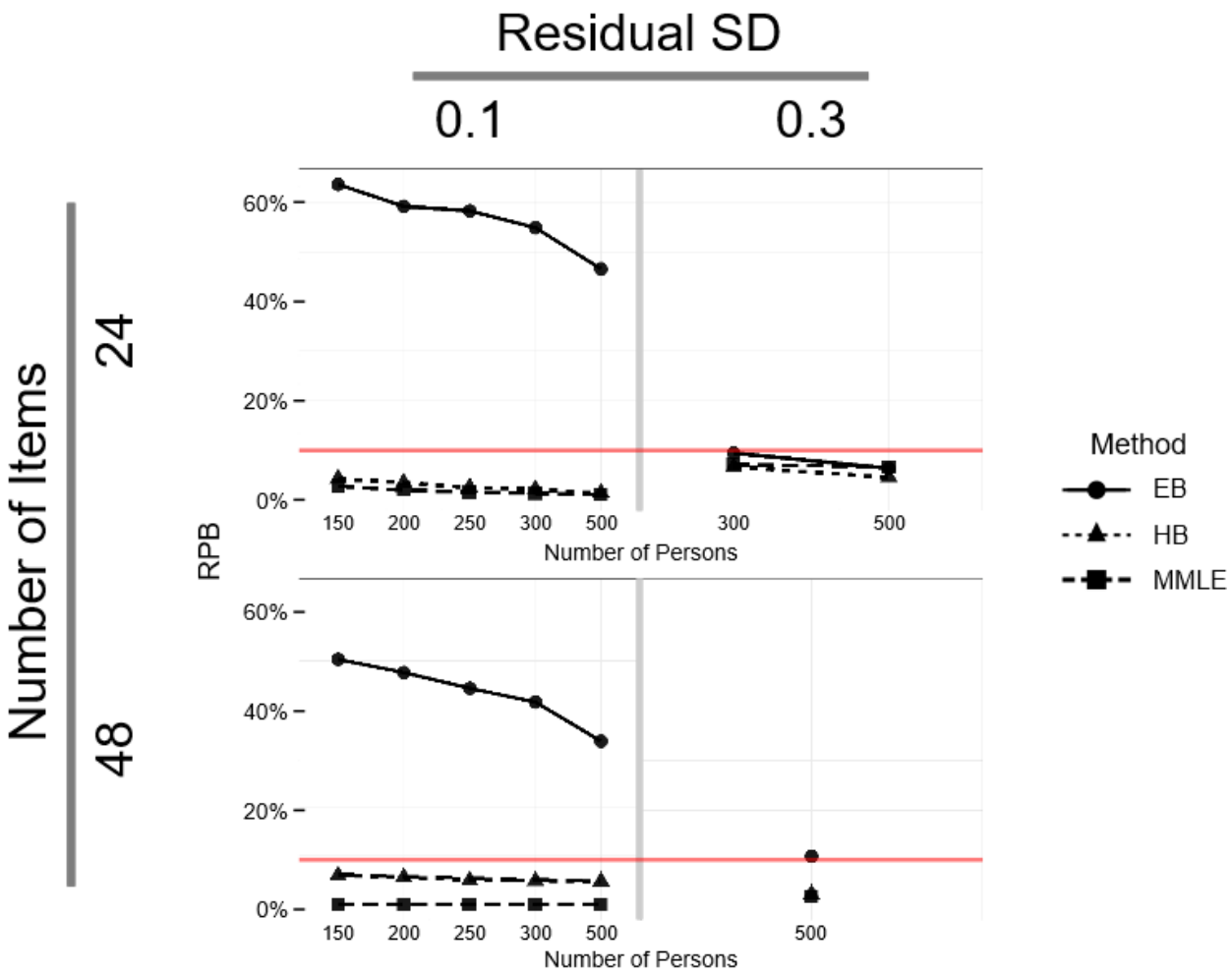


Figure 3: RPB comparison for MMLE, empirical Bayes, and hierarchical Bayes with item covariates

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

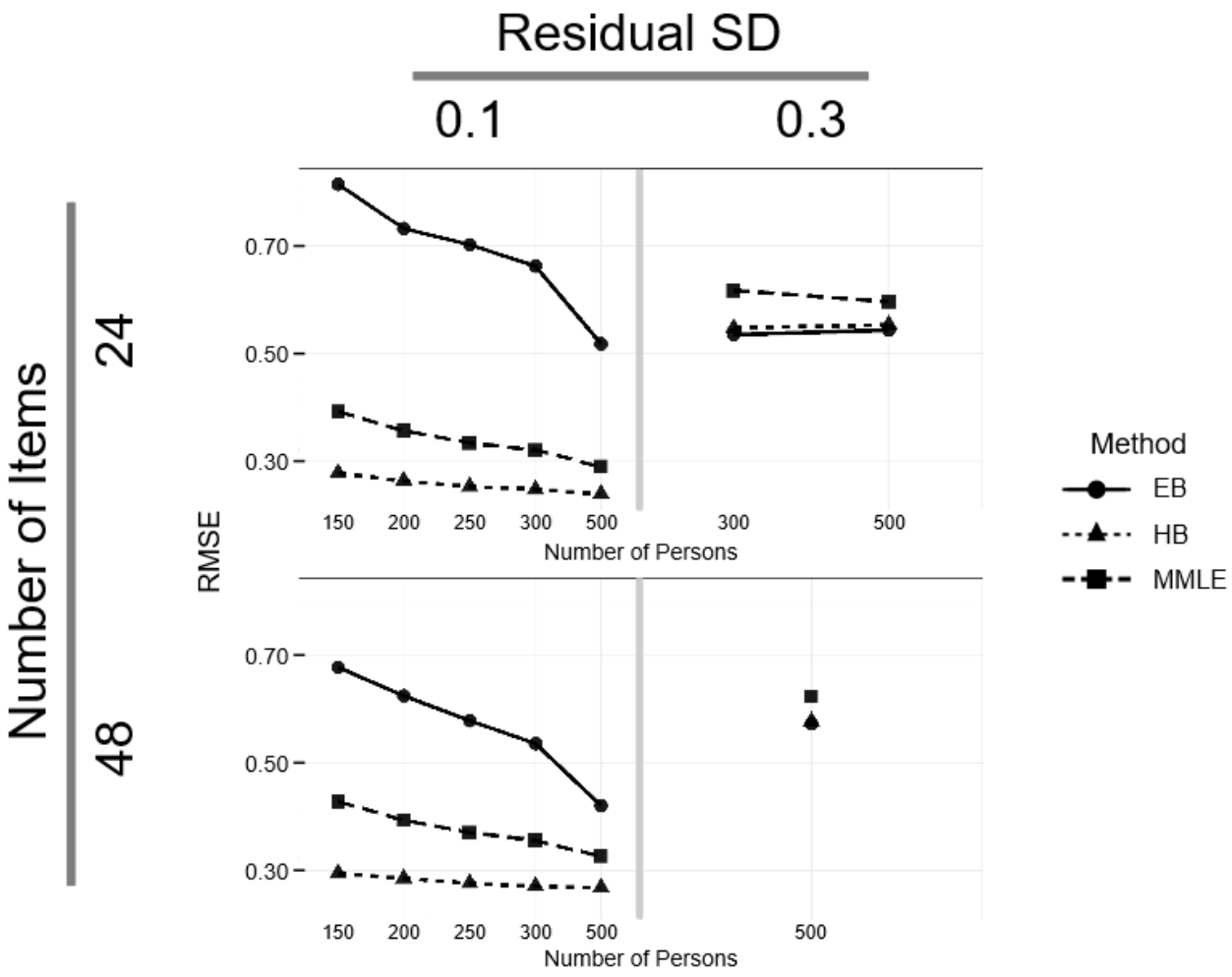


Figure 4: RMSE comparison for MMLE, empirical Bayes, and hierarchical Bayes with item covariates

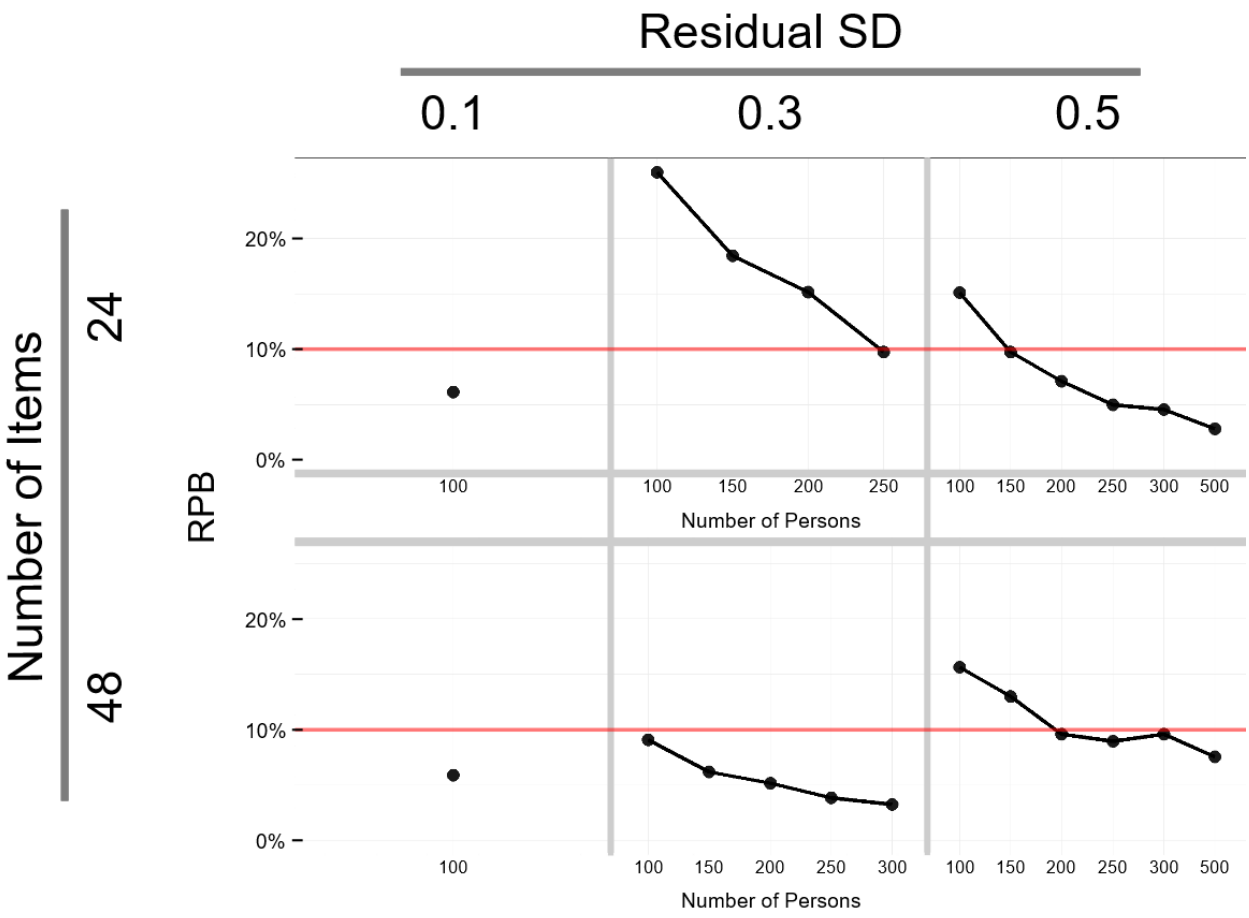


Figure 5: RPB for hierarchical Bayes with item covariates

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

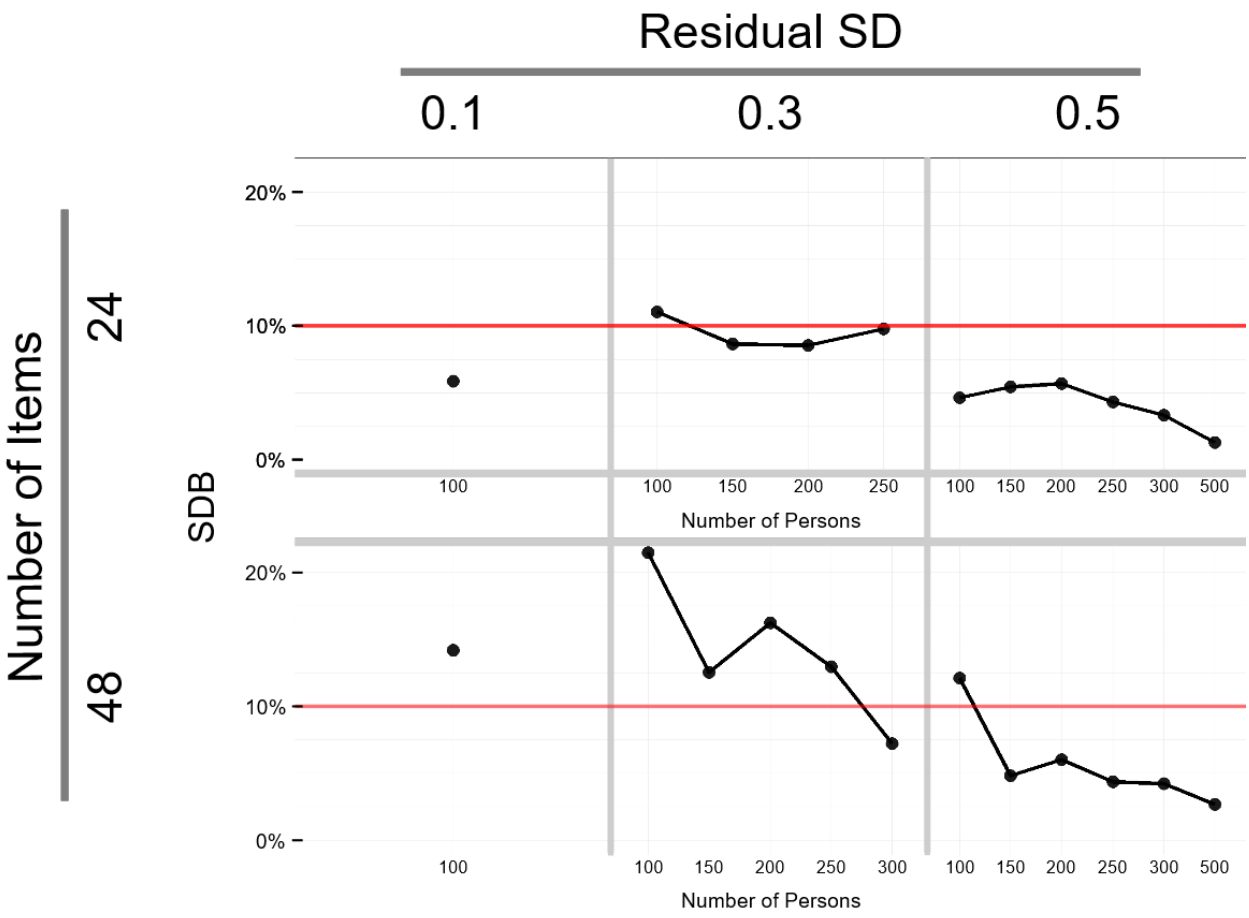


Figure 6: SDB for hierarchical Bayes with item covariates

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

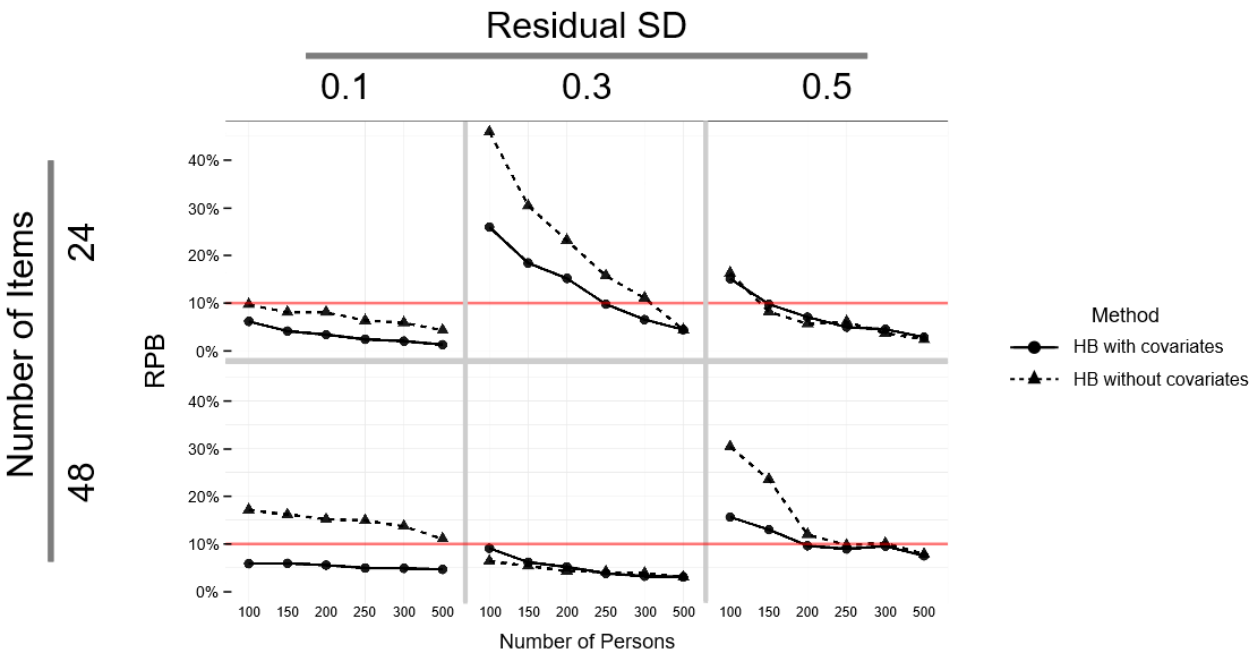


Figure 7: RPB comparison for hierarchical Bayes with covariates and hierarchical Bayes without covariates

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

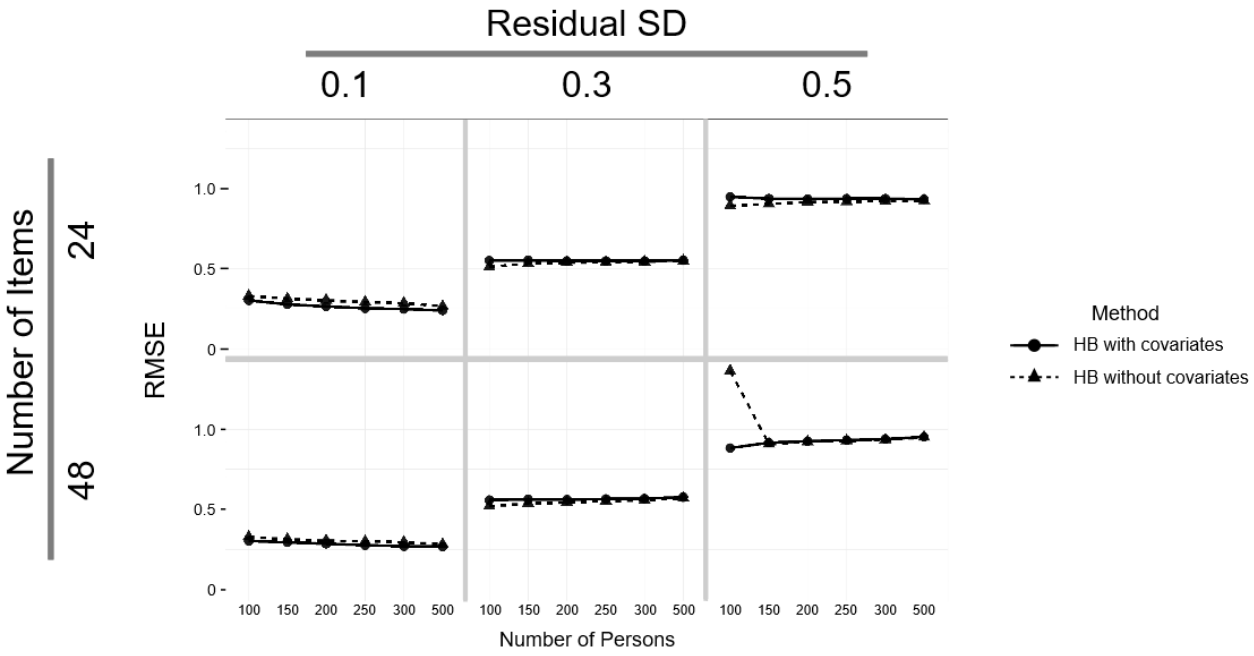


Figure 8: RMSE comparison for hierarchical Bayes with covariates and hierarchical Bayes without covariates

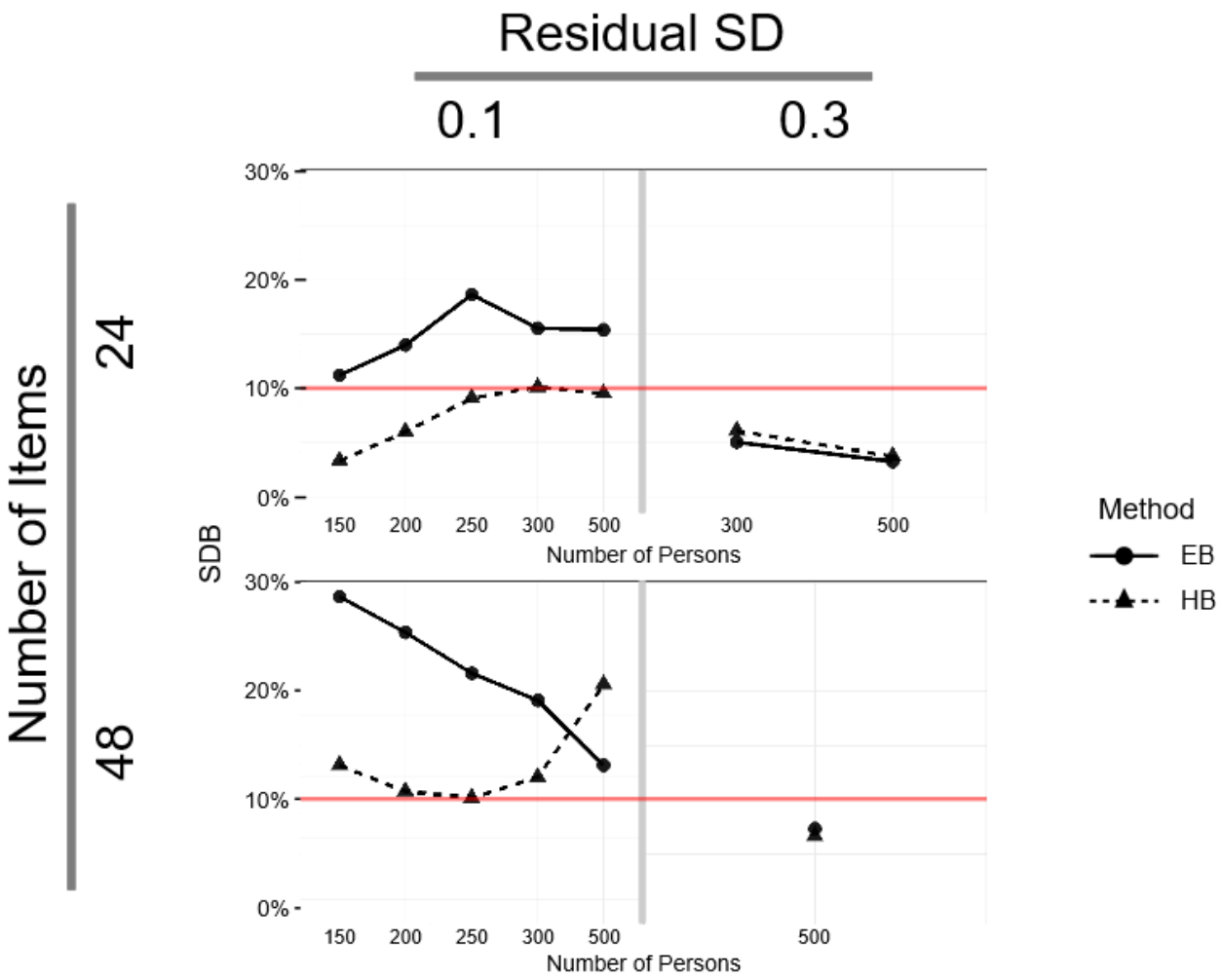


Figure 9: SDB comparison for empirical Bayes and hierarchical Bayes with covariates

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

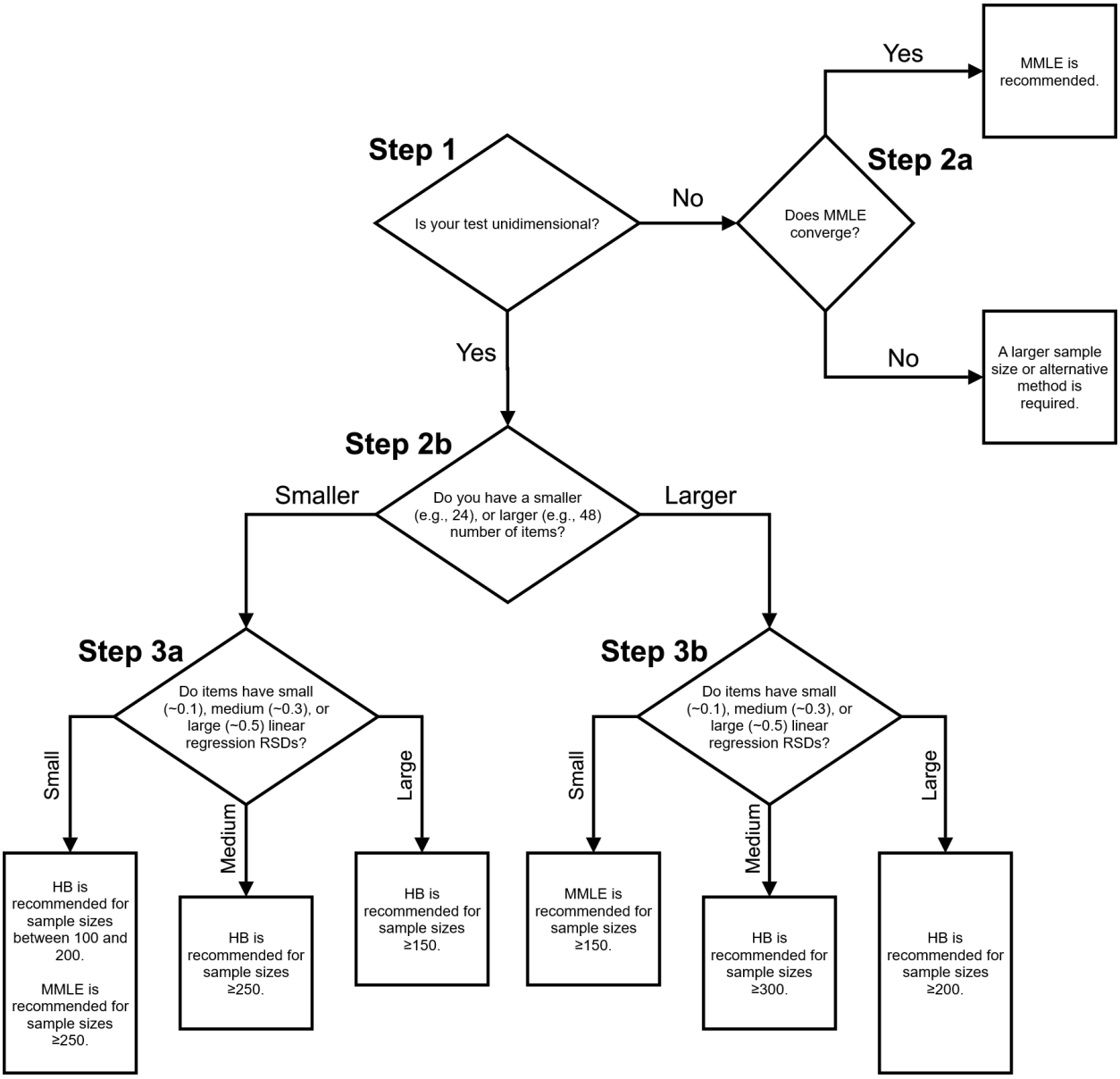


Figure 10: Method selection guideline

Appendix A

Investigation of Univariate vs. Multivariate Priors

To test whether using different priors on the item thresholds affects the posterior distributions of item parameters, sets of item parameters were generated under simulation conditions that would be the most sensitive to the effect of threshold covariances: 24 items, RSD = 0.5, and sample sizes of 100 and 250. For each of these two simulation conditions, three item parameter sets were generated: item parameters with uncorrelated item thresholds ($r = 0$), weakly correlated item thresholds ($r = 0.3$), and strongly correlated item thresholds ($r = 0.7$). This resulted in six item parameter sets (one for each combination of sample size and correlation strength). For each item parameter set, two hierarchical Bayes methods were used to estimate item parameters: one using a multivariate prior on item thresholds, and one using a univariate prior on item thresholds.

The only set of item parameters that resulted in varied results between the multivariate and univariate hierarchical Bayes methods was the item parameter set with a sample size of 100 and strongly correlated item thresholds ($r = 0.7$). With this set of item parameters, the RPBs of item parameters (averaged across items) were generally larger with the multivariate prior than with the univariate prior (except in the case of β_{i1}, where the reverse was true). Results were not highly comparable between the two methods in this condition, with correlations of item parameter estimates between the two methods ranging from $r = 0.046$ for α_i to $r = 0.660$ for β_{i3}. However, results were highly comparable between the multivariate and univariate methods in the remaining five conditions, with similar RPB (averaged across items) of item parameters and highly correlated item parameter estimates (having $r > 0.990$ for each item parameter between the two methods in the other five conditions).

Based on these results, we chose to use a univariate prior for all simulation conditions, for both generating and estimating item parameters, as used in the empirical Bayes method of the present study and in other Bayesian IRT estimation studies for polytomous item response models (e.g., Curtis, 2010; Fox, 2010; Kang et al., 2009). However, it is notable to mention

that in a situation where a researcher attempts to apply these methods when dealing with a small sample size (e.g., 100) and items with highly correlated thresholds (e.g., $r = 0.7$), results may vary depending on whether a univariate or multivariate prior is used.

Appendix B

R Code for Empirical Study and RSD Estimation with Classical Item Discrimination and Thresholds

The following code was designed to obtain MMLE, empirical Bayes, and hierarchical Bayes results.

```
################################################
###### Empirical Bayes estimation function: #####
################################################

EB_estimation_function <- function(EB_results_function_input){
Y <- HB_results_function_input[[1]]
variables <- HB_results_function_input[[2]]
I <- variables[[1]]
K <- variables[[2]]
D <- variables[[3]]
J <- variables[[4]]

ipg <- items_per_group <- I/(D+1)
item_group_index <- c()
for (group in 1:(D+1)){
item_group_index <- c(item_group_index, rep(group,ipg))
}
covariates <- matrix(nrow=I,ncol=D)
for (d in 1:D){
covariates[,d] <- c(rep(0,ipg*d),rep(1,ipg),rep(0,ipg*(D-d)))
}

library(mirt)
colnames_Y <- c()
for (item in 1:I){
colnames_Y <- c(colnames_Y, paste("Item",item,sep=" "))
}
colnames(Y) <- colnames_Y
mirt_Y <- mirt(Y, 1, itemtype='graded', method="EM", SE=TRUE)
summary_mirt <- coef(mirt_Y, printSE=TRUE, as.data.frame=TRUE)

if (dim(summary_mirt)[1]==(I*K + 2)){
mirt_estimates <- matrix(nrow=I,ncol=K)
mirt_SE <- matrix(nrow=I,ncol=K)
for (item in 1:I){
```

```
for (parameter in 1:K){
mirt_estimates[item,parameter] <- summary_mirt[((item-1)*K)+parameter,1]
mirt_SE[item,parameter] <- summary_mirt[((item-1)*K)+parameter,2]
}}
mirt_estimates[,2:K] <- -mirt_estimates[,2:K]

EB_regression_estimates <- matrix(nrow=(D+1),ncol=K)
EB_regression_SE <- c()
for (parameter in 1:K){
regression_structure <- cbind(mirt_estimates[,parameter],covariates)
colnames(regression_structure) <- c("Parameter",paste("Covariate",seq(1,D),sep='_'))
regression <- lm(as.formula(paste(colnames(regression_structure)[1],
paste(c(1, colnames(regression_structure)[2:(D+1)]), collapse=" + "), sep=" ~ ")),
data = data.frame(regression_structure))
summary_regression <- summary(regression)
EB_regression_estimates[,parameter] <- unname(summary_regression$"coefficients"[,1])
regression_residue <- unname(c(resid(regression)))
EB_regression_SE[parameter] <- sqrt(sum((regression_residue)^2) / (I - (D+1)))
}

EB_estimates <- matrix(nrow=I,ncol=K)
EB_SE <- matrix(nrow=I,ncol=K)
for (item in 1:I){
for (parameter in 1:K){
EB_estimates[item,parameter] <- (mirt_estimates[item,parameter]*(mirt_SE[item,parameter]^-2) +
EB_regression_estimates[item_group_index[item],parameter]*
(EB_regression_SE[parameter]^-2)) / (mirt_SE[item,parameter]^-2 + EB_regression_SE[parameter]^-2)
EB_SE[item,parameter] <- sqrt(1/(mirt_SE[item,parameter]^-2 + EB_regression_SE[parameter]^-2))
}}
return(list(mirt_estimates, mirt_SE, EB_regression_estimates, EB_regression_SE, EB_estimates, EB_SE))
}
else {
return(list("Error, MIRT unable to estimate parameters.","Error, MIRT unable to estimate parameters.",
"Error, MIRT unable to estimate parameters.","Error, MIRT unable to estimate parameters.",
"Error, MIRT unable to estimate parameters.","Error, MIRT unable to estimate parameters."))
}}

##################################################
###### Hierarchical Bayes estimation function: #####
##################################################
```

```
HB_estimation_function <- function(HB_results_function_input){
Y <- HB_results_function_input[[1]]
variables <- HB_results_function_input[[2]]
I <- variables[[1]]
K <- variables[[2]]
D <- variables[[3]]
J <- variables[[4]]

ipg <- items_per_group <- I/(D+1)
item_group_index <- c()
for (group in 1:(D+1)){
item_group_index <- c(item_group_index, rep(group,ipg))
}

covariates <- matrix(nrow=I,ncol=D)
for (d in 1:D){
covariates[,d] <- c(rep(0,ipg*d),rep(1,ipg),rep(0,ipg*(D-d)))
}

library(rstan)
Y_long <- c()
person_long <- c()
item_long <- c()
a <- 0
for (person in 1:J){
for (item in 1:I){
a <- a + 1
Y_long[a] <- Y[person,item] + 1  # Puts responses on a scale of 1-5 instead of 0-4 for Stan.
person_long[a] <- person
item_long[a] <- item
}}
data_stan <- list(number_categories=K, number_persons=J, number_items=I,
Y=Y_long, person=person_long, item=item_long, number_responses=length(Y_long), X=covariates, number_covariates=D)

GRM_stan <- "
data{
int<lower=0> number_categories;
int<lower=0> number_persons;
int<lower=0> number_items;
int<lower=0> number_responses;
int<lower=1, upper=number_categories> Y[number_responses];
```

```
int<lower=1, upper=number_persons> person[number_responses];
int<lower=1, upper=number_items> item[number_responses];
int<lower=0> number_covariates;
int<lower=0, upper=1> X[number_items, number_covariates];
}

parameters{
vector[number_persons] theta; //latent variable
real<lower=0> alpha[number_items]; //item discrimination
ordered[number_categories - 1] beta[number_items];  //category difficulty

vector[1 + number_covariates] gamma_alpha;
real<lower=0> phi_alpha;
vector[1 + number_covariates] gamma_beta[number_categories - 1];
real<lower=0> phi_beta[number_categories - 1];
}

model{

for (j in 1:number_persons){
theta[j] ~ normal(0,1);
}

phi_alpha ~ cauchy(0,10);

for (k in 1:number_categories-1){
phi_beta[k] ~ cauchy(0,10);
}

for (i in 1:number_items){
alpha[i] ~ normal((gamma_alpha[1] + gamma_alpha[2]*X[i,1] +
gamma_alpha[3]*X[i,2] + gamma_alpha[4]*X[i,3] + gamma_alpha[5]*X[i,4] + gamma_alpha[6]*X[i,5]), phi_alpha);
}

for (i in 1:number_items){
for (k in 1:(number_categories-1)){
beta[i,k] ~ normal((gamma_beta[k,1] + gamma_beta[k,2]*X[i,1] + gamma_beta[k,3]*X[i,2] +
gamma_beta[k,4]*X[i,3] + gamma_beta[k,5]*X[i,4] + gamma_beta[k,6]*X[i,5]), phi_beta[k]);
}}

for (c in 1:(number_covariates+1)){
```

```
gamma_alpha[c] ~ normal(0,10);
for (k in 1:(number_categories-1)){
gamma_beta[k,c] ~ normal(0,10);
}}

for (a in 1:number_responses){  // For each data point
  Y[a] ~ ordered_logistic(theta[person[a]]*alpha[item[a]], beta[item[a]]);
  }
}
"
fit_stan <- stan(model_code = GRM_stan, data = data_stan, seed=12212017)
summary_stan <- summary(fit_stan)
summary_stan <- summary_stan$summary

HB_estimates <- matrix(nrow=I,ncol=K)
HB_SD <- matrix(nrow=I,ncol=K)
HB_regression_estimates <- matrix(nrow=(D+1),ncol=K)
HB_regression_SD <- c()

start_index <- J + (I*K) + 1
end_index <- start_index + D
HB_regression_estimates[,1] <- summary_stan[start_index:end_index, 6]
HB_regression_SD[1] <- summary_stan[(end_index+1),6]
start_index <- J + (I*K) + D + 3
end_index <- start_index + (D+1)*(K-1) - 1
HB_regression_estimates[,2:K] <- matrix(summary_stan[start_index:end_index, 6],ncol=(K-1),byrow=FALSE)
HB_regression_SD[2:K] <- summary_stan[((end_index + 1):(end_index + K - 1)),6]

for (item in 1:I){
HB_estimates[item,1] <- summary_stan[J+item,6]
HB_SD[item,1] <- summary_stan[J+item,3]
for (threshold in 2:K){
HB_estimates[item,threshold] <- summary_stan[(J+I+(K-1)*(item-1)+(threshold-1)),6]
HB_SD[item,threshold] <- summary_stan[(J+I+(K-1)*(item-1)+(threshold-1)),3]
}}

return(list(HB_estimates, HB_SD, HB_regression_estimates, HB_regression_SD, summary_stan))
}

#########################################################################
###### Obtaining empirical Bayes & hierarchical Bayes estimates  #########
```

```
#########################################################################

data <- read.table(data.txt, sep="\t")
number_items <- 24
number_categories <- 5
number_covariates <- 5
number_persons <- 150
variables <- c(number_items, number_categories, number_covariates, number_persons)

estimation_input <- list(data, variables)

EB_estimation_output <- EB_estimation_function(estimation_input)
# Step 1: EB_estimation_output[[1]]: Maximum likelihood estimates
#         EB_estimation_output[[2]]: Maximum likelihood SEs
# Step 2: EB_estimation_output[[3]]: Regression estimates
#         EB_estimation_output[[3]]: Regression SEs
# Step 3: EB_estimation_output[[5]]: Empirical Bayes estimates
#         EB_estimation_output[[6]]: Empirical Bayes SEs

HB_estimation_output <- HB_estimation_function(estimation_input)
# HB_estimation_output[[1]]: Hierarchical Bayes estimates
# HB_estimation_output[[2]]: Hierarchical Bayes SDs
# HB_estimation_output[[3]]: Hierarchical Bayes RSD estimates
# HB_estimation_output[[4]]: Full Stan output
```

The following code uses classical item discrimination and thresholds to obtain the linear regression RSD estimates necessary to utilize the method selection guideline provided in Figure 10.

```
#######################################
###### RSD estimation function #########
#######################################

RSD_function <- function(RSD_function_input){
library(lme4)
library(psychometric)
RSD <- c()
Y <- RSD_function_input[[1]]
condition_variables <- RSD_function_input[[2]]
I <- condition_variables[[1]]
```

```
K <- condition_variables[[2]]
D <- condition_variables[[3]]
J <- condition_variables[[5]]

# If you have a unique item covariate structure, substitute D (above)
# with the correct covariate structure, and remove the following code chunk.
covariates <- matrix(0,nrow=I,ncol=D)
items_per_group <- I/(D+1)
covariates <- matrix(nrow=I,ncol=D)
for (covariate in 1:D){
covariates[,covariate] <- c(rep(0,items_per_group*covariate),rep(1,items_per_group),
rep(0,items_per_group*(D-covariate)))
}

# Obtains item discriminations.
discrimination <- unname(unlist(item.exam(Y, discrim=TRUE)$Discrimination))

parameter_types <- matrix(0,nrow=I*K,ncol=K-1)
# Creates a covariate matrix for all item parameter types,
# to be used with the linear regression later on in this function.
for (parameter in 2:K){
parameter_types[((parameter-1)*I+1):(parameter*I),parameter-1] <- 1
}

for (threshold in 1:(K-1)){
Y_threshold <- Y
Y_threshold[Y_threshold < threshold] <- 0
Y_threshold[Y_threshold >= threshold] <- 1
item_means <- colMeans(Y_threshold)
difficulty <- c()
parameter_estimates <- discrimination
for (item in 1:I){
difficulty[item] <- min(4,max(-4, -qlogis(item_means[item], scale=discrimination[item])))
# Obtains each item threshold, with a minimum of -4 and maximum of +4.
}
parameter_estimates <- c(parameter_estimate, difficulty)
}

regression_data <- data.frame(parameter_estimates, parameter_types, covariates)
# Combines item parameter esimates, parameter covariates, and item covariates for running regression.
regression_data_names <- c("Estimate")
```

```
for (parameter in 1:(K-1)){
regression_data_names <- c(regression_data_names, paste("P",parameter,sep=''))
}
for (d in 1:D){
regression_data_names <- c(regression_data_names, paste("X",d,sep=''))
}
names(regression_data) <- regression_data_names
regression_equation <- "Estimate ~ 1"
for (parameter in 1:(K-1)){
regression_equation <- paste(regression_equation, paste("+ P",parameter,sep=''),sep=' ')
}
for (covariate in 1:D){
regression_equation <- paste(regression_equation, paste("+ X",covariate,sep=''), sep=' ')
}

regression <- lm(formula = as.formula(regression_equation), data = regression_data)
# Runs linear regression on CTT item parameter estimates.
regression_summary <- summary(regression)
RSD <- regression_summary$sigma
return(RSD)
}

########################################
####### Obtaining RSD estimate #########
########################################

setwd("C:\\Users\\myname\\DataLocation") # Replace this with the directory where the data is located
data <- read.table("data.txt",sep="\t") # Replace "data.txt" with the name of the data file

number_items <- 24
number_categories <- 5
number_covariates <- 5
number_persons <- 100
condition_variables <- list(number_items, number_categories, number_covariates, number_persons)
coefficients <- c(-0.2703644864, 0.8808019178, -0.0001940328, 0.0001785366)
# Based on our simulation study data.

CTT_RSD <- RSD_function(list(data, condition_variables))
RSD_estimate <- coefficients[1] + coefficients[2]*CTT_RSD +
coefficients[3]*number_items + coefficients[4]*number_persons # Final result.
```

Appendix C

Supplemental Empirical Study Results

Table C1: *Regression Coefficients for Empirical Bayes vs. Hierarchical Bayes, J=273*

	Empirical Bayes		Hierarchical Bayes				
	EST	SE	Mean	Median	SD	0.025*	0.975*
Discrimination							
$\gamma_{\alpha 0}$	2.192	0.254	2.191	2.188	0.278	1.663	2.754
$\gamma_{\alpha 1}$	0.637	0.370	0.492	0.493	0.370	-0.239	1.205
$\gamma_{\alpha 2}$	1.018	0.360	1.042	1.049	0.375	0.302	1.803
$\gamma_{\alpha 3}$	0.493	0.360	0.399	0.397	0.374	-0.357	1.132
ϕ_{α}	0.805	0.151	0.781	0.768	0.112	0.601	1.037
Threshold 1							
$\gamma_{\beta 01}$	-4.269	0.558	-4.171	-4.168	0.539	-5.216	-3.107
$\gamma_{\beta 11}$	-1.944	0.811	-2.079	-2.059	0.788	-3.641	-0.568
$\gamma_{\beta 21}$	-1.405	0.789	-1.402	-1.396	0.761	-2.876	0.080
$\gamma_{\beta 31}$	0.265	0.789	0.479	0.466	0.742	-1.029	1.923
$\phi_{\beta 1}$	1.764	0.726	1.650	1.625	0.229	1.263	2.174
Threshold 2							
$\gamma_{\beta 02}$	-2.385	0.435	-2.265	-2.257	0.434	-3.138	-1.421
$\gamma_{\beta 12}$	-1.799	0.632	-1.855	-1.849	0.605	-3.063	-0.679
$\gamma_{\beta 22}$	-1.067	0.615	-1.039	-1.041	0.598	-2.229	0.157
$\gamma_{\beta 32}$	0.538	0.615	0.659	0.653	0.593	-0.503	1.826
$\phi_{\beta 2}$	1.375	0.441	1.302	1.288	0.168	1.015	1.663
Threshold 3							
$\gamma_{\beta 03}$	-0.458	0.352	-0.352	-0.359	0.387	-1.115	0.431
$\gamma_{\beta 13}$	-1.208	0.511	-1.375	-1.375	0.528	-2.401	-0.323
$\gamma_{\beta 23}$	-0.491	0.497	-0.444	-0.441	0.511	-1.413	0.590
$\gamma_{\beta 33}$	0.738	0.497	0.787	0.784	0.512	-0.203	1.776
$\phi_{\beta 3}$	1.112	0.288	1.133	1.120	0.148	0.881	1.470
Threshold 4							
$\gamma_{\beta 04}$	1.480	0.297	1.578	1.584	0.356	0.894	2.269
$\gamma_{\beta 14}$	-0.449	0.432	-0.688	-0.695	0.478	-1.638	0.234
$\gamma_{\beta 24}$	0.002	0.420	0.054	0.045	0.471	-0.856	1.000
$\gamma_{\beta 34}$	0.948	0.420	0.923	0.924	0.479	-0.040	1.840
$\phi_{\beta 4}$	0.940	0.206	1.043	1.031	0.136	0.816	1.352

Note. * Percentiles of posterior distribution

Table C2: *Comparison of Empirical and Hierarchical Bayesian Estimates for α_i with J=273*

		Empirical							Hierarchical	
		Step 1		Step 2		Step 3				
Item	Domain	$\widehat{\alpha}_i$	$\widehat{\tau}_{\alpha i}$	$\bar{\alpha}_i$	$\bar{\phi}_\alpha$	$\tilde{\alpha}_i$	$\tilde{\sigma}_{\alpha i}$	Shrinkage	Posterior Median	SD
1	C	1.395	0.211	NA	NA	NA	NA	NA	1.372	0.203
2	C	2.766	0.277	2.829	0.805	2.773	0.262	0.106	2.723	0.258
3	C	4.115	0.392	2.829	0.805	3.868	0.352	0.192	3.868	0.337
4	C	2.124	0.240	2.829	0.805	2.181	0.230	0.082	2.092	0.222
5	C	3.148	0.311	2.829	0.805	3.107	0.290	0.130	3.055	0.278
6	C	3.122	0.295	2.829	0.805	3.087	0.277	0.119	3.128	0.273
7	C	3.012	0.276	2.829	0.805	2.993	0.261	0.105	3.047	0.266
8	C	2.256	0.226	2.829	0.805	2.298	0.218	0.073	2.326	0.223
9	C	2.726	0.268	2.829	0.805	2.736	0.254	0.100	2.780	0.256
10	C	2.187	0.212	2.829	0.805	2.228	0.205	0.065	2.293	0.220
11	E	1.405	0.169	3.210	0.805	1.481	0.166	0.042	1.519	0.176
12	E	3.264	0.306	3.210	0.805	3.257	0.286	0.127	3.254	0.290
13	E	2.399	0.232	3.210	0.805	2.461	0.223	0.077	2.494	0.234
14	E	3.647	0.345	3.210	0.805	3.580	0.317	0.155	3.629	0.318
15	E	3.858	0.358	3.210	0.805	3.751	0.327	0.165	3.757	0.325
16	E	2.941	0.271	3.210	0.805	2.968	0.257	0.102	3.029	0.279
17	E	3.511	0.329	3.210	0.805	3.468	0.305	0.144	3.502	0.303
18	E	3.948	0.369	3.210	0.805	3.820	0.335	0.174	3.853	0.337
19	E	3.767	0.358	3.210	0.805	3.675	0.327	0.165	3.767	0.333
20	E	3.362	0.316	3.210	0.805	3.341	0.294	0.134	3.393	0.303
21	P	4.125	0.392	2.685	0.805	3.849	0.352	0.192	3.651	0.319
22	P	2.607	0.265	2.685	0.805	2.615	0.252	0.098	2.498	0.240
23	P	2.558	0.242	2.685	0.805	2.568	0.232	0.083	2.613	0.238
24	P	4.201	0.397	2.685	0.805	3.903	0.356	0.196	3.750	0.325
25	P	4.209	0.390	2.685	0.805	3.919	0.351	0.190	3.910	0.334
26	P	2.050	0.200	2.685	0.805	2.087	0.194	0.058	2.091	0.197
27	P	1.213	0.147	2.685	0.805	1.260	0.144	0.032	1.289	0.149
28	P	2.696	0.249	2.685	0.805	2.695	0.238	0.088	2.722	0.244
29	P	1.284	0.151	2.685	0.805	1.332	0.148	0.034	1.349	0.154
30	P	1.904	0.190	2.685	0.805	1.945	0.185	0.053	1.929	0.187
31	S	2.147	0.218	2.192	0.805	2.150	0.210	0.068	2.129	0.203
32	S	1.958	0.197	2.192	0.805	1.971	0.191	0.056	2.012	0.198
33	S	3.049	0.290	2.192	0.805	2.951	0.273	0.115	2.904	0.266
34	S	2.264	0.239	2.192	0.805	2.259	0.229	0.081	2.193	0.222
35	S	1.714	0.177	2.192	0.805	1.736	0.173	0.046	1.761	0.178
36	S	2.228	0.215	2.192	0.805	2.226	0.208	0.067	2.283	0.214
37	S	2.958	0.269	2.192	0.805	2.881	0.255	0.101	2.885	0.258
38	S	1.897	0.190	2.192	0.805	1.912	0.185	0.053	1.953	0.188
39	S	2.144	0.205	2.192	0.805	2.147	0.199	0.061	2.164	0.199
40	S	1.560	0.173	2.192	0.805	1.588	0.169	0.044	1.629	0.176

Note. Maximum likelihood estimation was unable to estimate the fourth threshold of item 1 ($\beta_{1,4}$) because no responses were obtained in the fifth category of item 1. Because of this, item 1 was omitted during Step 2 when obtaining regression estimates, resulting in the missing values for item 1 (noted as NA).

Table C3: *Comparison of Empirical and Hierarchical Bayesian Estimates for $\beta_{i,4}$ with J=273*

		Empirical							Hierarchical	
		Step 1		Step 2		Step 3				
Item	Domain	$\widehat{\beta}_{i,4}$	$\widehat{\tau}_{\beta i,4}$	$\bar{\beta}_{i,4}$	$\bar{\phi}_{\beta,4}$	$\tilde{\beta}_{i,4}$	$\tilde{\sigma}_{\beta i,4}$	Shrinkage	Posterior Median	SD
1	C	NA	NA	NA	NA	NA	NA	NA	-1.463	0.194
2	C	0.466	0.217	1.031	0.940	0.494	0.211	0.051	0.606	0.228
3	C	1.550	0.297	1.031	0.940	1.503	0.283	0.091	1.635	0.296
4	C	-0.538	0.193	1.031	0.940	-0.474	0.189	0.040	-0.391	0.196
5	C	0.351	0.234	1.031	0.940	0.391	0.227	0.059	0.500	0.244
6	C	1.486	0.249	1.031	0.940	1.456	0.241	0.066	1.597	0.260
7	C	2.679	0.285	1.031	0.940	2.540	0.273	0.084	2.758	0.293
8	C	1.117	0.203	1.031	0.940	1.113	0.199	0.045	1.231	0.215
9	C	0.930	0.220	1.031	0.940	0.935	0.214	0.052	1.069	0.230
10	C	1.240	0.200	1.031	0.940	1.231	0.195	0.043	1.346	0.217
11	E	0.065	0.155	1.482	0.940	0.102	0.153	0.027	0.163	0.168
12	E	2.055	0.272	1.482	0.940	2.011	0.261	0.077	2.186	0.275
13	E	1.640	0.223	1.482	0.940	1.632	0.217	0.053	1.786	0.240
14	E	1.076	0.263	1.482	0.940	1.106	0.254	0.073	1.255	0.275
15	E	1.856	0.295	1.482	0.940	1.822	0.281	0.090	1.978	0.304
16	E	2.571	0.276	1.482	0.940	2.485	0.264	0.079	2.699	0.281
17	E	1.539	0.269	1.482	0.940	1.535	0.258	0.076	1.690	0.285
18	E	1.830	0.298	1.482	0.940	1.799	0.284	0.091	1.963	0.306
19	E	0.929	0.269	1.482	0.940	0.971	0.258	0.076	1.131	0.284
20	E	1.262	0.254	1.482	0.940	1.277	0.245	0.068	1.438	0.272
21	P	2.159	0.318	2.428	0.940	2.187	0.302	0.103	2.160	0.304
22	P	-0.116	0.208	2.428	0.940	0.002	0.203	0.047	0.093	0.209
23	P	1.978	0.235	2.428	0.940	2.005	0.228	0.059	2.140	0.246
24	P	2.775	0.344	2.428	0.940	2.734	0.323	0.118	2.716	0.323
25	P	2.896	0.348	2.428	0.940	2.839	0.327	0.121	2.913	0.336
26	P	3.390	0.293	2.428	0.940	3.304	0.279	0.088	3.454	0.283
27	P	2.112	0.200	2.428	0.940	2.126	0.195	0.043	2.214	0.205
28	P	2.969	0.286	2.428	0.940	2.923	0.274	0.085	3.082	0.288
29	P	2.559	0.228	2.428	0.940	2.552	0.221	0.055	2.636	0.227
30	P	3.553	0.307	2.428	0.940	3.445	0.292	0.097	3.580	0.304
31	S	0.793	0.191	1.480	0.940	0.820	0.187	0.040	0.915	0.198
32	S	1.042	0.186	1.480	0.940	1.058	0.183	0.038	1.160	0.195
33	S	1.053	0.235	1.480	0.940	1.078	0.228	0.059	1.174	0.241
34	S	-0.505	0.199	1.480	0.940	-0.420	0.194	0.043	-0.314	0.198
35	S	1.941	0.205	1.480	0.940	1.920	0.201	0.046	2.031	0.214
36	S	1.736	0.215	1.480	0.940	1.723	0.209	0.050	1.855	0.224
37	S	3.214	0.308	1.480	0.940	3.046	0.293	0.097	3.208	0.295
38	S	1.998	0.211	1.480	0.940	1.973	0.206	0.048	2.087	0.216
39	S	2.736	0.256	1.480	0.940	2.649	0.247	0.069	2.789	0.248
40	S	0.791	0.166	1.480	0.940	0.812	0.163	0.030	0.904	0.173

Note. Maximum likelihood estimation was unable to estimate the fourth threshold of item 1 ($\beta_{1,4}$) because no responses were obtained in the fifth category of item 1. Because of this, item 1 was omitted during Step 2 when obtaining regression estimates, resulting in the missing values for item 1 (noted as NA).

Appendix D

Supplemental Simulation Study Results

Table D1: *MMLE Convergence Rate for 500 Replications*

		Number of Persons					
Number of Items	Residual SD	100	150	200	250	300	500
24	0.1	99.8%	100%	100%	100%	100%	100%
24	0.3	91.8%	97.8%	99.4%	99.8%	100%	100%
24	0.5	16.0%	42.4%	54.2%	72.0%	83.6%	95.8%
48	0.1	99.8%	100%	100%	100%	100%	100%
48	0.3	51.0%	78.2%	90.2%	96.8%	98.8%	100%
48	0.5	0.40%	6.60%	12.20%	26.0%	35.4%	69.6%

Table D2: *RPB for MMLE by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	1.709	1.517	1.107	2.739	2.557
24	0.1	200	1.340	1.872	1.764	1.034	1.417
24	0.1	250	0.795	1.154	0.823	1.417	1.385
24	0.1	300	0.583	1.222	1.059	0.805	1.063
24	0.1	500	0.375	0.944	0.947	0.709	0.829
24	0.3	300	0.547	5.030	3.866	7.129	5.872
24	0.3	500	0.352	5.033	4.538	6.476	5.333
48	0.1	150	1.271	5.817	5.464	5.927	6.117
48	0.1	200	1.081	5.246	4.916	5.646	5.751
48	0.1	250	0.572	5.338	5.443	4.480	5.060
48	0.1	300	0.679	5.044	4.805	4.890	5.045
48	0.1	500	0.358	4.588	4.542	4.721	4.753
48	0.3	500	0.358	-2.312	-2.609	-1.744	-1.855

Table D3: *RPB for Empirical Bayes by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	-34.020	-39.749	-63.628	-13.104	-19.702
24	0.1	200	-33.123	-35.476	-59.218	-13.033	-18.852
24	0.1	250	-32.417	-34.127	-58.262	-11.881	-17.625
24	0.1	300	-31.597	-32.036	-54.943	-11.426	-16.432
24	0.1	500	-27.868	-24.751	-46.576	-8.977	-12.675
24	0.3	300	-9.344	-0.777	-7.945	-9.042	1.369
24	0.3	500	-6.248	1.371	-3.008	-4.473	2.441
48	0.1	150	-31.468	-32.854	-50.321	-10.012	-16.339
48	0.1	200	-30.479	-30.136	-47.551	-9.209	-15.130
48	0.1	250	-29.961	-27.633	-44.373	-9.043	-13.884
48	0.1	300	-28.804	-25.362	-41.550	-8.149	-12.764
48	0.1	500	-25.295	-19.178	-33.500	-5.830	-9.166
48	0.3	500	-7.131	-7.672	-10.690	-4.203	-5.382

Table D4: *RPB for Hierarchical Bayes with Item Covariates by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	4.119	0.894	0.597	2.031	1.741
24	0.1	200	3.393	1.311	1.463	0.569	0.822
24	0.1	250	2.392	0.692	0.633	0.933	0.812
24	0.1	300	2.007	0.900	1.139	0.246	0.500
24	0.1	500	1.281	0.641	0.881	0.334	0.430
24	0.3	300	2.918	3.824	-0.761	-6.570	4.575
24	0.3	500	1.727	4.310	1.762	-1.976	4.456
48	0.1	150	3.321	4.343	4.498	5.929	4.830
48	0.1	200	2.775	4.137	4.342	5.551	4.691
48	0.1	250	2.086	4.409	4.987	4.439	4.158
48	0.1	300	1.998	4.236	4.421	4.850	4.290
48	0.1	500	1.303	3.992	4.196	4.679	4.281
48	0.3	500	2.172	-3.058	-2.482	-2.538	-2.810

Table D5: *RMSE for MMLE by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	0.205	0.366	0.253	0.267	0.392
24	0.1	200	0.177	0.344	0.228	0.240	0.357
24	0.1	250	0.156	0.322	0.209	0.221	0.334
24	0.1	300	0.142	0.308	0.195	0.209	0.320
24	0.1	500	0.110	0.282	0.172	0.180	0.290
24	0.3	300	0.146	0.616	0.345	0.295	0.601
24	0.3	500	0.112	0.596	0.323	0.270	0.571
48	0.1	150	0.197	0.396	0.260	0.281	0.428
48	0.1	200	0.169	0.367	0.235	0.256	0.393
48	0.1	250	0.150	0.350	0.218	0.235	0.370
48	0.1	300	0.138	0.337	0.206	0.223	0.356
48	0.1	500	0.105	0.307	0.178	0.196	0.326
48	0.3	500	0.106	0.609	0.326	0.341	0.623

Table D6: *RMSE for Empirical Bayes by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	0.396	0.814	0.597	0.220	0.448
24	0.1	200	0.380	0.732	0.556	0.203	0.417
24	0.1	250	0.367	0.701	0.546	0.188	0.391
24	0.1	300	0.355	0.662	0.513	0.180	0.364
24	0.1	500	0.311	0.518	0.439	0.154	0.291
24	0.3	300	0.178	0.534	0.303	0.255	0.518
24	0.3	500	0.133	0.543	0.294	0.244	0.519
48	0.1	150	0.375	0.677	0.475	0.193	0.379
48	0.1	200	0.360	0.624	0.448	0.179	0.348
48	0.1	250	0.350	0.578	0.420	0.169	0.322
48	0.1	300	0.338	0.536	0.396	0.161	0.302
48	0.1	500	0.294	0.420	0.326	0.143	0.249
48	0.3	500	0.127	0.571	0.309	0.318	0.573

Table D7: *RMSE for Hierarchical Bayes by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	0.156	0.275	0.186	0.189	0.278
24	0.1	200	0.132	0.263	0.172	0.173	0.262
24	0.1	250	0.118	0.253	0.161	0.163	0.251
24	0.1	300	0.110	0.247	0.153	0.157	0.246
24	0.1	500	0.094	0.238	0.140	0.144	0.236
24	0.3	300	0.155	0.547	0.305	0.255	0.525
24	0.3	500	0.118	0.553	0.299	0.245	0.524
48	0.1	150	0.130	0.260	0.165	0.195	0.295
48	0.1	200	0.119	0.253	0.159	0.187	0.285
48	0.1	250	0.111	0.252	0.151	0.175	0.276
48	0.1	300	0.106	0.247	0.145	0.170	0.271
48	0.1	500	0.090	0.240	0.135	0.159	0.268
48	0.3	500	0.107	0.571	0.304	0.313	0.576

Table D8: *SDB for Empirical Bayes by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	-1.322	-6.472	-10.202	4.114	5.339
24	0.1	200	-2.223	-10.249	-12.988	3.996	6.155
24	0.1	250	-2.017	-10.192	-17.640	4.182	3.795
24	0.1	300	-3.682	-14.543	-13.297	3.543	3.830
24	0.1	500	-8.441	-6.633	-14.387	4.407	4.638
24	0.3	300	6.029	5.551	2.453	2.766	5.052
24	0.3	500	4.136	3.876	2.058	3.254	4.257
48	0.1	150	23.822	29.738	14.627	18.610	27.652
48	0.1	200	22.579	23.756	11.953	15.127	26.812
48	0.1	250	20.382	17.192	7.859	14.423	23.463
48	0.1	300	16.372	14.616	7.097	13.780	21.265
48	0.1	500	11.080	13.998	6.541	13.018	15.967
48	0.3	500	8.336	6.878	4.586	3.216	6.432

Table D9: *SDB for Hierarchical Bayes by Item Parameter Type*

# Items	RSD	# Persons	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}
24	0.1	150	-0.611	2.257	4.316	3.669	2.787
24	0.1	200	6.980	6.350	6.246	5.000	6.534
24	0.1	250	10.101	8.387	8.140	8.189	8.275
24	0.1	300	10.981	7.295	11.119	7.898	9.159
24	0.1	500	9.573	8.847	10.508	8.139	9.268
24	0.3	300	7.071	3.425	3.132	5.212	4.344
24	0.3	500	3.965	2.368	2.370	4.735	3.814
48	0.1	150	14.322	15.962	14.958	12.387	13.125
48	0.1	200	13.817	12.836	9.182	8.156	12.434
48	0.1	250	12.814	11.161	10.638	10.611	13.295
48	0.1	300	14.756	13.645	13.261	13.939	14.982
48	0.1	500	22.554	12.750	12.980	13.695	12.050
48	0.3	500	7.656	3.944	4.661	4.555	4.950

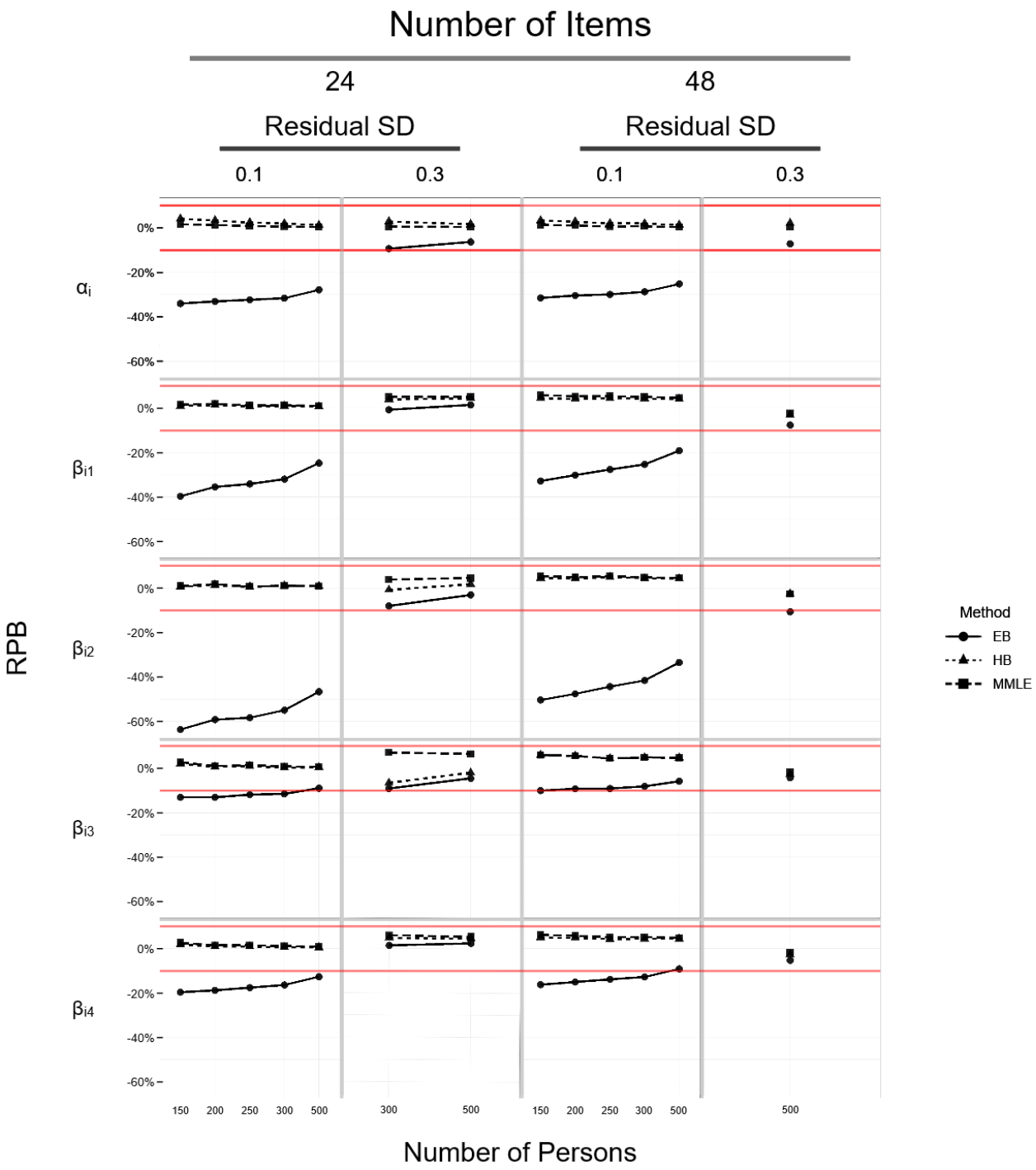


Figure D1: RPB comparison for MMLE, empirical Bayes, and hierarchical Bayes with item covariates, separated by item parameter type

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

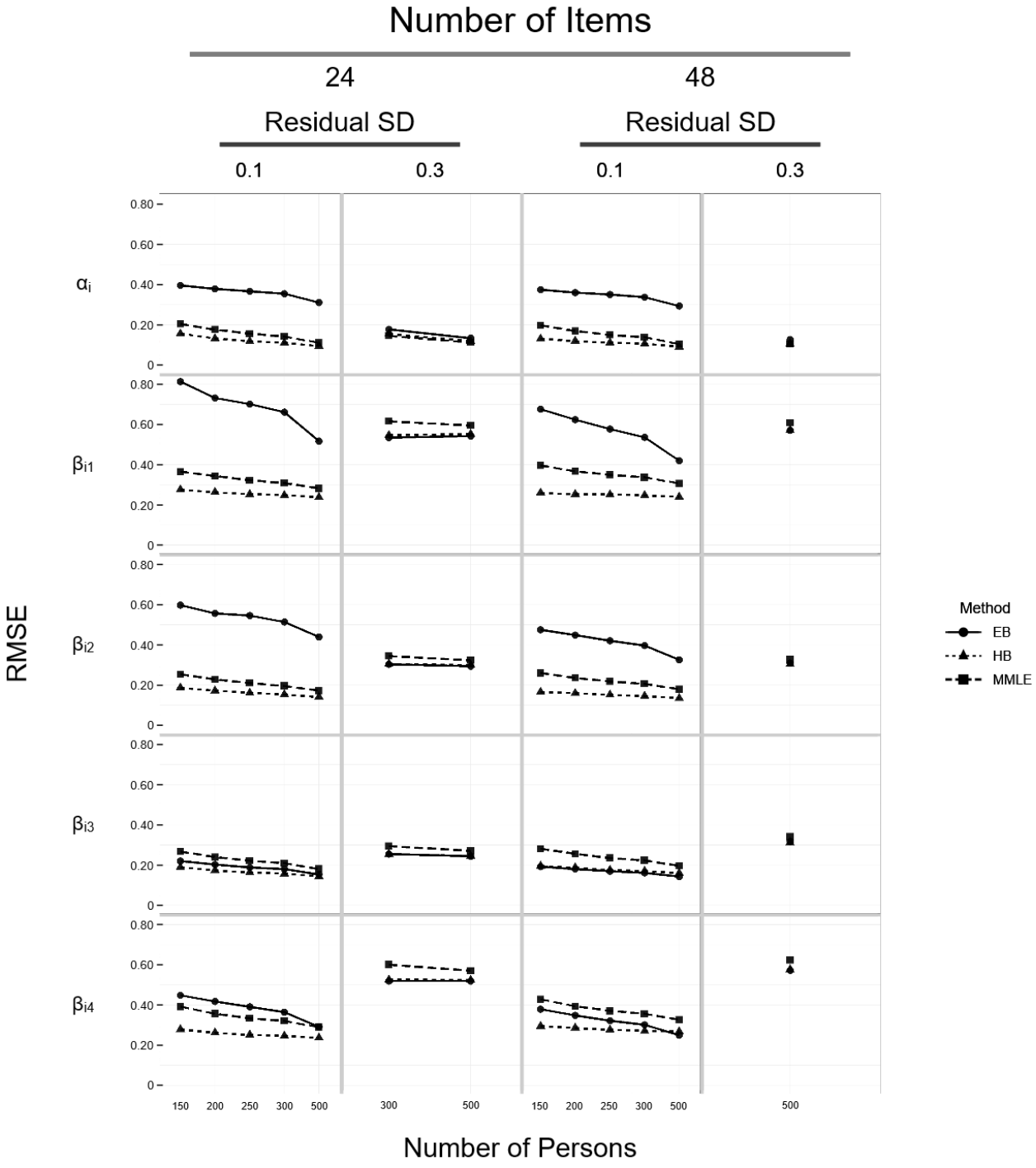


Figure D2: RMSE comparison for MMLE, empirical Bayes, and hierarchical Bayes with item covariates, separated by item parameter type

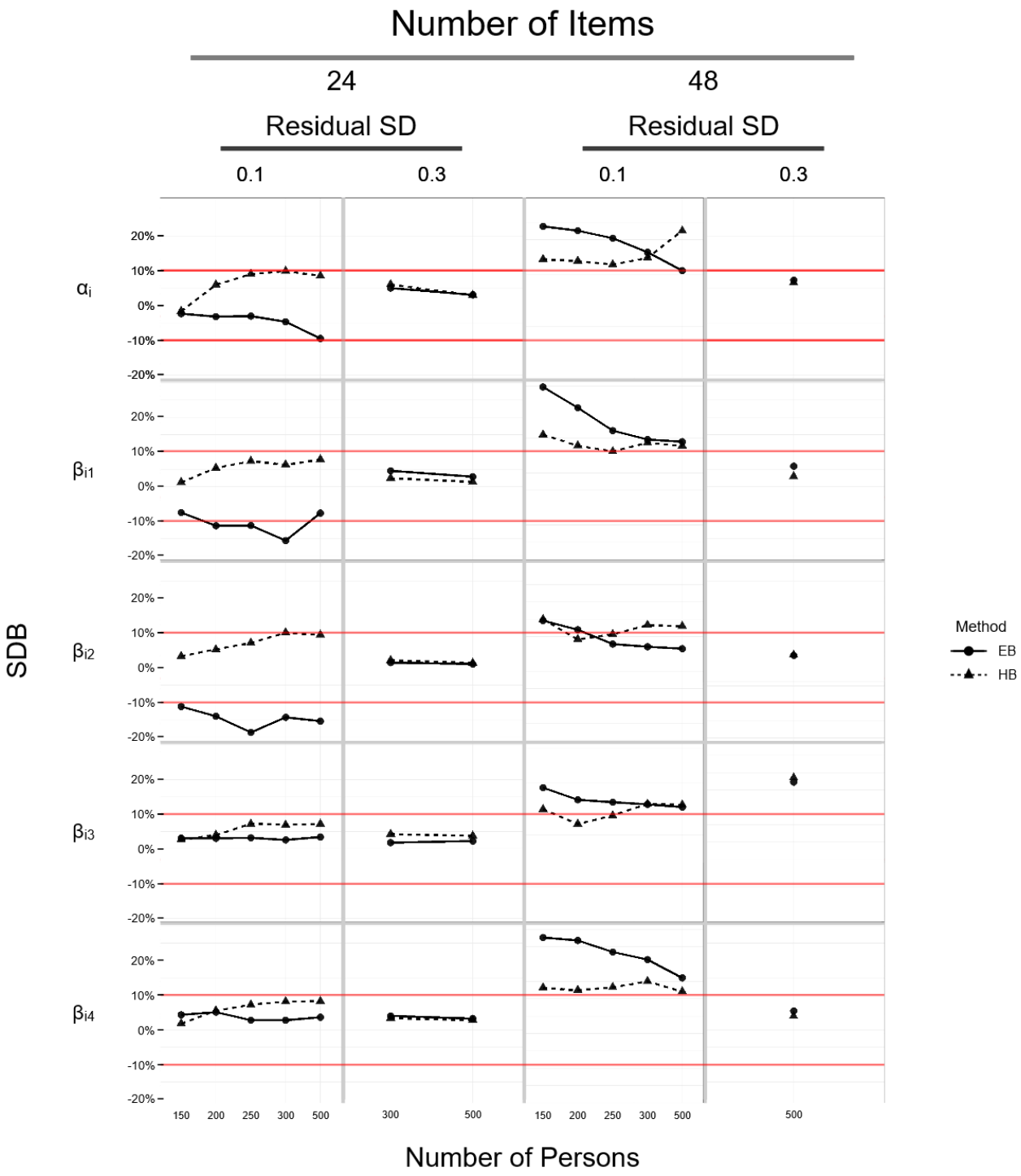


Figure D3: SDB comparison for MMLE, empirical Bayes, and hierarchical Bayes with item covariates, separated by item parameter type

Note. Horizontal lines indicate cutoff for acceptable RPB (10%).

Appendix E

Method Selection Guideline Supplement

Table E1: *Method Selection Guideline*

			Hierarchical Bayes			MMLE
Items	RSD	Persons	$RPB < 10\%$	$SDB < 10\%$	Acceptable?	100% Convergence?
24	0.1	100	Yes	Yes	Yes	No
24	0.1	150	Yes	Yes	Yes	Yes
24	0.1	200	Yes	Yes	Yes	Yes
24	0.1	250	Yes	No	No	Yes
24	0.1	300	Yes	No	No	Yes
24	0.1	500	Yes	No	No	Yes
24	0.3	100	No	No	No	No
24	0.3	150	No	Yes	No	No
24	0.3	200	No	Yes	No	No
24	0.3	250	Yes	Yes	Yes	No
24	0.3	300	Yes	Yes	Yes	Yes
24	0.3	500	Yes	Yes	Yes	Yes
24	0.5	100	No	Yes	No	No
24	0.5	150	Yes	Yes	Yes	No
24	0.5	200	Yes	Yes	Yes	No
24	0.5	250	Yes	Yes	Yes	No
24	0.5	300	Yes	Yes	Yes	No
24	0.5	500	Yes	Yes	Yes	No
48	0.1	100	Yes	No	No	No
48	0.1	150	Yes	No	No	Yes
48	0.1	200	Yes	No	No	Yes
48	0.1	250	Yes	No	No	Yes
48	0.1	300	Yes	No	No	No
48	0.1	500	Yes	No	No	Yes
48	0.3	100	Yes	No	No	No
48	0.3	150	Yes	No	No	No
48	0.3	200	Yes	No	No	No
48	0.3	250	Yes	No	No	No
48	0.3	300	Yes	Yes	Yes	No
48	0.3	500	Yes	Yes	Yes	Yes
48	0.5	100	No	No	No	No
48	0.5	150	No	Yes	No	No
48	0.5	200	Yes	Yes	Yes	No
48	0.5	250	Yes	Yes	Yes	No
48	0.5	300	Yes	Yes	Yes	No
48	0.5	500	Yes	Yes	Yes	No

Note. *Hierarchical Bayes was considered an acceptable method if both $RPB < 10\%$ and $SDB < 10\%$ for all item parameter types.
**MMLE was considered an acceptable method if (i) HB was not considered an acceptable method, and (ii) MMLE had 100% convergence for all 500 replications.