Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Adaptive Shrinkage with Student's t: Limited Gains over Gaussian Hierarchical Linear Models

G. Buhrman [*†], J.-S. Kim [†], and W. Lyu [‡]

†Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin, USA
‡Faculty of Education, University of Macau, Macau, China
*Corresponding author. Email: buhrman@wisc.edu

Abstract

Standard Hierarchical Linear Models (HLM) assume that unobserved random effects follow a Gaussian distribution, a premise central to the mechanism of partial pooling. However, when the true distribution of random effects is heavy-tailed, the resulting shrinkage may pull outlying observations too aggressively toward the mean, potentially biasing estimates for clusters in the tails. This study investigates "adaptive shrinkage" by implementing Bayesian HLM, explicitly specifying a Student's t distribution for the random intercepts to allow for greater flexibility in modeling extreme values. Through a comprehensive simulation study, the classification accuracy, robustness, and computational trade-offs of the standard Gaussian HLM, the Student's t approach, and a robust linear mixed-effects model were compared. Data-generating conditions varied true random intercept distributions (Student's t with df = 3, 5, 10; normal mixtures), intraclass correlation, and cluster sizes. Results indicate that while the Student's t distribution alters the allocation of shrinkage–specifically weakening its correlation with cluster size–it does not substantially improve the classification of clusters in the tails or reduce estimation error of these clusters compared to standard Gaussian HLM. Furthermore, the standard Gaussian HLM demonstrated remarkable robustness to normality violations, whereas the Student's t models introduced computational costs, including lower convergence rates and higher frequencies of divergent transitions. These findings suggest that simply relaxing the normality assumption via a Student's t distribution is insufficient for improving tail-specific inferences in HLM. Future research should explore alternative methods for adjusting the shrinkage weighting formula to better handle extreme clusters.

Keywords: multilevel modeling, longitudinal data analysis, random effects

1. Introduction

In research involving multilevel data, a common analytic goal is the decomposition and modeling of variability between clusters. These clustered data often take the form of students within schools, patients within hospitals, or items within a test. When employing hierarchical linear models (HLM) to account for this nested structure, the standard approach assumes that unobserved random effects follow a Gaussian distribution (Raudenbush & Bryk, 2002).

This assumption is central to the mechanism of partial pooling, which occupies the middle ground between no pooling (where each cluster is treated as an entirely independent group) and complete pooling (where all observations are treated as part of a single group). While it is standard practice to assume normally distributed random effects, this assumption is frequently violated in practice (Micceri, 1989). Such violations may occur because the true population distribution is heavy-tailed or multimodal (Verbeke & Lesaffre, 1996; Yuan et al., 2004).

To understand how violations of this assumption impact model estimates, it is helpful to first explicitly define the standard HLM framework. Consider a basic unconditional random-intercept model:

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij} = \beta_{0j} + e_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + u_{0j} \end{aligned} \tag{1}$$

where Y_{ij} is the outcome for individual i ($i = 1, \ldots, n_j$) nested within cluster j ($j = 1, \ldots, J$). At Level 1, β_{0j} represents the unobserved true mean for cluster j, and e_{ij} is the within-cluster residual, assumed to be normally distributed such that $e_{ij} \sim N(0, \sigma_e^2)$. At Level 2, γ_{00} is the grand mean across all clusters, and u_{0j} is the random intercept (the cluster-specific deviation from the grand mean), traditionally assumed to follow a Gaussian distribution, $u_{0j} \sim N(0, \sigma_u^2)$.

Violations of this normality assumption for u_{0j} become particularly significant due to the role of shrinkage. Because HLM utilizes partial pooling, the final estimate of a random effect depends on both cluster-specific information and the overall sample mean. The degree of shrinkage a cluster's random intercept experiences depends on its reliability factor, $\hat{\lambda}_j$, calculated as:

$$\hat{\lambda}_j = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_j}} \tag{2}$$

This reliability factor directly influences the partially pooled estimate of the cluster-specific effect, $\hat{\beta}_{0j}$:

$$\hat{\beta}_{0j} = \hat{\lambda}_j \bar{Y}_j + (1 - \hat{\lambda}_j)\hat{\gamma}_{00}. \tag{3}$$

Here, $\bar{Y}_j$ is the unpooled observed mean for cluster j. The estimated random effect is the deviation of the partially pooled estimate from the estimated grand mean:

$$\hat{u}_{0j} = \hat{\beta}_{0j} - \hat{\gamma}_{00}. \tag{4}$$

Consequently, the estimated random effect is simply the observed deviation between the cluster mean and the estimated grand mean, scaled by the reliability factor:

$$\hat{u}_{0j} = \hat{\lambda}_j(\bar{Y}_j - \hat{\gamma}_{00}). \tag{5}$$

Because σ_u^2 is typically estimated under a Gaussian assumption, the model assumes extreme values are improbable and applies pressure to pull potential outliers back toward the mean. For this reason, this study evaluates whether using a Student's t distribution for the random effects in an HLM improves the classification of clusters in the tails and the estimation of these clusters' effects relative to the standard Gaussian HLM.

1.1 Motivating Problem

The Gaussian assumption is computationally convenient but problematic when the true distribution is heavy-tailed or multimodal. In such cases, shrinkage may pull observations in the tails toward the center more aggressively than necessary, biasing the estimates of outlying clusters. This is a particular concern when the goal of modeling is to rank-order clusters or determine whether a cluster falls above or below a specific cutoff threshold (Arling et al., 2007; Arpino & Varriale, 2010; Herrmann et al., 2016).

Using data from the High School and Beyond Study of 1982, it can be seen that clusters identified as being in the top or bottom 10% vary depending on whether the slope was estimated using no

pooling or partial pooling (see Figure 1). The shrinkage that occurs under partial pooling can change whether a cluster is classified into the top or bottom tail or in the center of the distribution. Therefore, clusters whose slopes were near the threshold were likely reclassified based on the use of pooling and subsequent shrinkage. It is necessary to have accurate classification of schools in this context for a number of reasons, the first among them being accountability. In this context, the slope of socioeconomic status (SES) in different schools is estimated, meaning the goal is to determine if students will have an even playing field in a school based on their home resource level or family income. Ideally, it is desirable for schools to have a small SES slope, meaning that students achieve the same outcomes at this school regardless of how many resources they have. Identifying schools that provide atypical advantages or disadvantages based on socioeconomic status is an essential component of evaluating school equity and accountability.

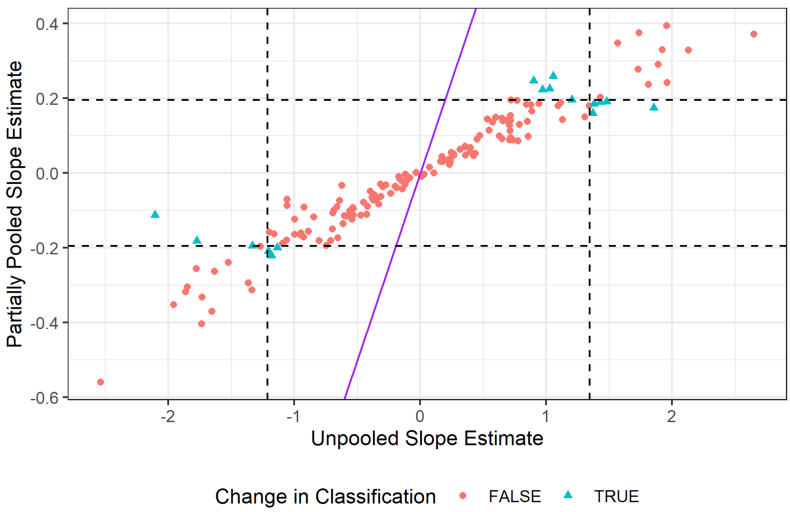

Figure 1. High School and Beyond 1982: Changes in classification of clusters based on SES random slopes

2. Adaptive Shrinkage

Recently, there has been a renewed discussion about the impact of normality assumption violations in HLM (Hui et al., 2021). Previous research suggests that using the Student's *t* distribution instead of a Gaussian distribution could help address violations of the normality assumption, particularly when the true distribution is heavy-tailed (Pinheiro et al., 2001; Tong & Zhang, 2012). A potential benefit of using the Student's *t* distribution is that the model may more accurately classify extreme clusters in the tails.

By altering the assumptions under which σ_u^2 is estimated, it may be possible to adjust the "dial" on shrinkage–a concept referred to as adaptive shrinkage. Adaptive shrinkage can be further defined as any step taken to alter the effective shrinkage weights. In a standard Gaussian HLM, the reliability formula relies on a single, estimated global variance parameter ($\hat{\sigma}_u^2$) applied uniformly to all clusters. This means that shrinkage is driven almost entirely by cluster size. By assuming the random effects follow a Student's *t* distribution, the principle of the reliability formula remains the same, but the implementation of this random effects assumption changes the operational formula. Understanding how adaptive shrinkage is implemented is necessary to understanding how the operational formula for reliability changes.

In this study, adaptive shrinkage is implemented via Bayesian estimation. Because a Hamiltonian

Monte Carlo (HMC) framework is utilized, group-level random effects are explicitly sampled as latent parameters. Therefore, assigning a Student's t prior to these parameters is functionally equivalent to specifying a Student's t distribution for the random intercepts, aligning with the parameterizations seen in earlier robust HLM literature (Pinheiro et al., 2001; Tong & Zhang, 2012). This distribution was specified directly in the model formula using the `brms` grouping syntax `(1 | gr(group_id, dist = "student"))`. To ensure efficient sampling geometry within the HMC framework, `brms` internally utilizes a data augmentation approach. Specifically, it parameterizes the Student's t distribution as a scale mixture of normal distributions by estimating a latent, cluster-specific variance multiplier drawn from an inverse chi-square distribution. This latent, cluster-specific variance multiplier allows the variance to localize rather than relying solely on the estimate of σ_u^2. This localization of variance allows the estimated reliability, and thus the effective shrinkage weight, to adapt to extreme observations independently of sample size.

In the current study, the focus is solely on random intercepts. This study compares using a Student's t distribution for the random intercepts to the standard Gaussian HLM and a robust HLM approach designed to mitigate outlier influence at all levels (Koller, 2016). Unlike conventional HLM, which minimizes the squared errors and thus magnifies the impact of extreme values, the robust approach uses bounded weight functions to reduce the weight of outliers and prevent them from disrupting the estimation of variance components. The estimands of primary concern in this study are the posterior means of the random intercepts. These random intercepts are then used to determine whether a cluster should be in either the top or bottom 10% of all clusters.

3. Simulation Study

A simulation study was conducted to answer the following questions: (1) given a truly heavy-tailed distribution of random intercepts, which modeling approaches result in the most accurate classification of clusters in the tails of said distribution? (2) in terms of classification of clusters in the tails, how robust is the standard Gaussian HLM to a violation of the normality assumption? and (3) what are the trade-offs that come from using adaptive shrinkage compared to the typical Gaussian distribution?

All data were simulated and analyzed using R Statistical Software (R Core Team, 2025). Implementation of the Bayesian estimation of random intercepts was performed using the `brms` package (Bürkner, 2017). Implementation of the robust linear mixed-effects model for estimation was performed using the `robustlmm` package (Koller, 2016).

3.1 Data and Manipulated Factors

The model used to generate the data for the simulation was a conditional random intercepts model, which can be defined as

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij};\ e_{ij} \sim N(0,1) \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ & \beta_{1j} = \gamma_{01} \end{aligned} \tag{6}$$

where individuals $i = 1, \ldots, n_j$ are nested in groups $j = 1, \ldots, J$, X_{ij} is an individual-level predictor, and W_j is a cluster-level predictor. The true fixed effects were set to constants typical of standardized coefficients in applied research ($\gamma_{00} = 0$, $\gamma_{10} = 0.3$, $\gamma_{01} = 0.1$).

The core manipulation in the simulation design was to change the distribution used to generate the random intercepts, u_{0j}, to a heavy-tailed distribution or a mixture of normal distributions. Both a Student's t distribution and a mixture of normal distributions were used to generate heavy-tailed random effects. For the Student's t distribution, the degrees of freedom parameter was varied to be either 3, 5, or 10. The core manipulation is explicitly defined in equation (7).

$$\begin{aligned} &\text{Heavy-Tailed:} && u_{0j} \sim t\left(\nu, 0, \frac{\sigma_u^2(\nu-2)}{\nu}\right) \\ &\text{Mixture-Normal:} && u_{0j} \sim 0.9 \times N(0, \sigma^2_{\text{narrow}}) + 0.1 \times N(0, \sigma^2_{\text{wide}}) \end{aligned} \tag{7}$$

In (7), ν is the degrees of freedom, σ_u^2 is the variance of the random intercepts. The σ^2_{wide} and σ^2_{narrow} terms are determined based on the input parameter σ_u^2 and are calculated according to equation (8).

$$\begin{aligned} \kappa &= \frac{1}{(0.10 \times 3) + (0.90 \times 0.5)} \\ \sigma^2_{\text{narrow}} &= \sigma_u^2 \times 0.5\kappa \\ \sigma^2_{\text{wide}} &= \sigma_u^2 \times 3\kappa \end{aligned} \tag{8}$$

A critical feature of this simulation was the standardization of variance across all conditions. By scaling non-normal distributions to match the population variance of a hypothetical normal distribution, the impact of the distribution's shape (heavy-tailedness) was isolated from its spread (variance).

Other manipulated factors included the intraclass correlation coefficient (0.05 or 0.25), the number of clusters (30, 50, or 100), the average cluster size, $\bar{n}$, (10 or 30), and the variation in cluster size, α, (0 or 0.50). The levels of these factors were chosen to reflect the common ranges of these values in applied multilevel studies. For specifics on how cluster size was varied, see equation (9).

$$n_j \sim \text{Unif}[(1-\alpha)\bar{n}, (1+\alpha)\bar{n}] \tag{9}$$

Table 1 contains a summary of all of the simulation conditions we examined in this study.

Table 1. Simulation Conditions

Factor	Levels
True Random Intercept Distribution	Student's t (df = 3, 5, 10), Normal Mixture
Intraclass Correlation Coefficient (ICC)	0.05, 0.25
Number of Clusters (J)	30, 50, 100
Average Cluster Size ($\bar{n}$)	10, 30
Cluster Size Variation (α)	0, 0.50
Number of Replications	200

Two hundred replications of data were generated for each unique simulation condition, defined by crossing the manipulation factors in a design grid. For all iterations, the tails were defined as the bottom and top 10% of clusters.

To quantify the performance of the various estimation approaches, several criteria were employed. First, Brier scores were used to determine how accurately each approach predicted a cluster's membership in the tails of the distribution. Second, the correlation between group size and the magnitude of shrinkage experienced by a random intercept prediction was calculated—a quantity hereafter referred to as "shrinkage correlation". Third, the coverage of the random intercepts was assessed, defined as the proportion of iterations in which the true simulated random intercept fell within a 95% credible interval. Finally, the root mean squared error (RMSE) was calculated specifically for random intercepts located in the tails of the distribution. All performance criteria are formally defined in Table 2.

Table 2. Performance Criteria Formulae

Performance Criteria	Formula
Brier Score	$\frac{1}{J}\sum_{j=1}^{J}(p_j - o_j)^2$
Shrinkage Correlation	$\frac{\sum_{j=1}^{J}(n_j-\bar{n})(s_j-\bar{s})}{\sqrt{\sum_{j=1}^{J}(n_j-\bar{n})^2\sum_{j=1}^{J}(s_j-\bar{s})^2}}$
Coverage of Random Intercepts	$\frac{1}{J}\sum_{j=1}^{J} I(\hat{u}_{0j,L} \leq u_{0j} \leq \hat{u}_{0j,U})$
RMSE of Tails	$\sqrt{\frac{1}{J_{tails}}\sum_{j\in\text{tails}}(\hat{u}_{0j} - u_{0j})^2}$

1 For the Brier score, p_j refers to probability of the random intercept of cluster j being in the tails, and o_j refers to whether or not the random intercept for cluster j is truly in the tails.
2 For the adaptive shrinkage, s_j represents the shrinkage experienced by the random intercept for cluster j and $\bar{s}$ is the average amount of shrinkage experienced by the random intercepts for all clusters.
3 For the coverage, $\hat{u}_{0j,L}$ refers to the 2.5th percentile of the posterior distribution for u_{0j} and $\hat{u}_{0j,U}$ refers to the 97.5th percentile of the posterior distribution for u_{0j}.

3.2 Bayesian Estimation Specifications

All Bayesian models were estimated using the `brms` package via the `cmdstanr` backend. Four parallel Markov chains were run for each model, with each chain running for a total of 4,000 iterations, where 2,000 of these iterations were treated as burn-in. This resulted in 8,000 posterior draws per model. To ensure rigorous sampling and minimize divergent transitions, the target average proposal acceptance probability was set to 0.995 and the maximum tree depth was set to 15. Model convergence was formally evaluated using the Gelman-Rubin $\hat{R}$ statistic (Gelman & Rubin, 1992; Vehtari et al., 2021). Convergence was considered successful if $\hat{R} < 1.05$ for all estimated parameters. Additionally, models were continuously monitored for divergent transitions post-burn-in to ensure the validity and reliability of the posterior distributions.

Weakly informative priors were strategically utilized to stabilize estimation. For the population-level intercept and residual standard deviation, the default `brms` prior was used, which is a Student's t distribution with 3 degrees of freedom, a location of 0, and a scale of 2.5. For the predictors, fixed effects were assigned Normal(0, 1) priors. This constitutes a weakly informative prior, as the true values of the fixed effects fall well within the high-density region of the Normal(0,1) distribution.

To stabilize estimation and encourage the Student's t models to converge, the random effect standard deviation was assigned a Half-Normal(0, 0.5) prior. While the standard Gaussian HLM could easily converge without this specification, this same Half-Normal(0, 0.5) prior was intentionally used on the random effect standard deviation of the standard Gaussian HLM as well. This choice was made to ensure that any observed differences in shrinkage mechanics were driven entirely by the specified distribution of the random effects, rather than being confounded by different levels of prior regularization on the variance components.

Estimated models included both models that used a fixed value for the degrees of freedom (df) parameter and models where the df parameter was estimated from the data. In conditions where the df parameter was fixed, it was set to either 3, 5, or 10 depending on the true value of the df used to generate the data. In conditions where the df parameter was estimated from the data, it was assigned a Gamma(3, 0.5) prior with a lower bound of 2 to ensure finite variance.

3.3 Results

Though the type of heavy-tailed distribution and the true value of the df were varied in the data-generation process, the results were qualitatively similar across the df = 3, 5, 10, and mixture of normals conditions. For the sake of visual clarity, only the results from the df = 10 data generation

conditions are presented.

Regarding the classification accuracy of clusters based on their random intercepts–specifically whether a cluster fell within the top or bottom 10% of the population–the choice of estimation model made little difference (see Figure 2).

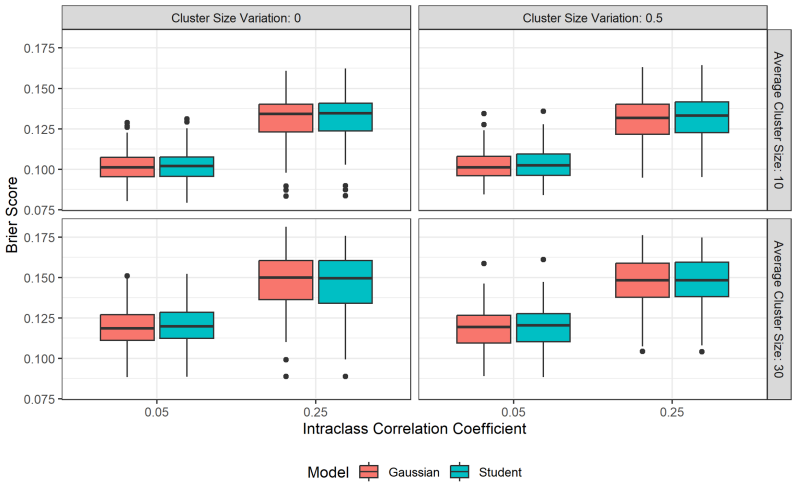

Figure 2. Brier scores for data generated using Student's t (df = 10)

Similarly, the estimation model had a negligible impact on the bias and variance of the random intercepts estimated within the tails of the population distribution (see Figure 3).

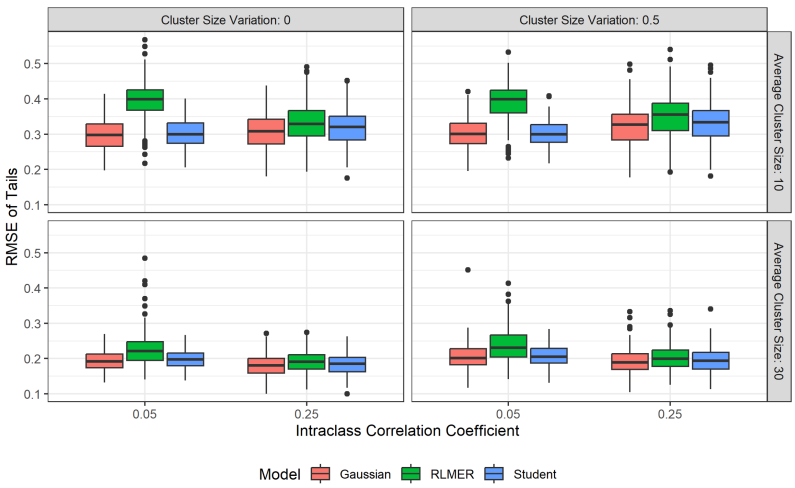

Figure 3. RMSE of tails for data generated using Student's t (df = 10)

Furthermore, random intercept coverage did not differ significantly between the standard Gaussian HLM and the model utilizing a Student's t distribution. In contrast, the robust linear mixed-effects model yielded substantially lower coverage rates (see Figure 4).

Variations were observed, however, in the correlation between shrinkage and sample size depending on the model used to estimate the random intercepts (see Figure 5).

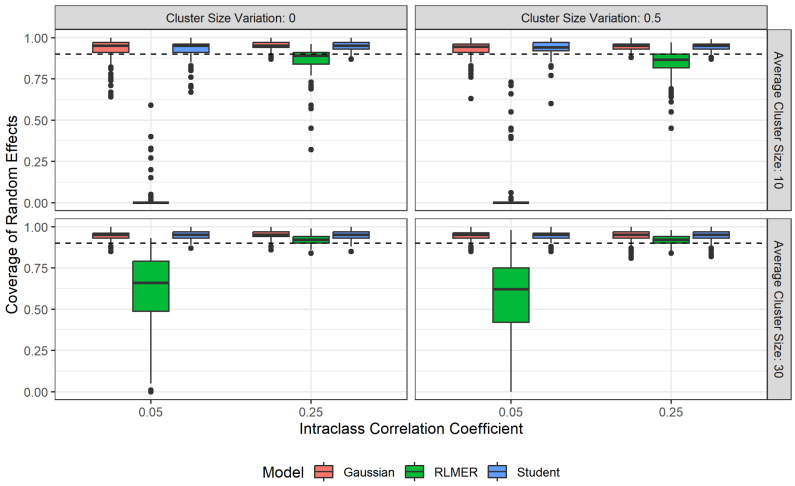

Figure 4. Coverage of random intercepts using data generated from Student's t (df = 10)

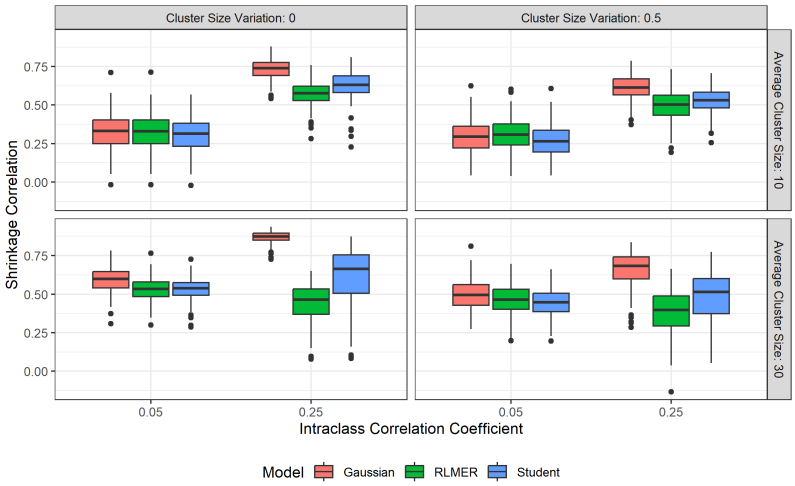

Figure 5. Shrinkage correlations for data generated from Student's t (df = 10)

Convergence rates and the presence of divergent transitions were also recorded across all simulation iterations. The standard Gaussian HLM and the Student's *t* model with a fixed degrees-of-freedom hyperparameter both achieved 100% convergence. However, when the degrees-of-freedom hyperparameter was estimated as part of the model-fitting process with a $Gamma(\alpha = 3, \theta = 2)$ prior on df, convergence rates for the Student's *t* models—applied to data generated from Student's *t* distributions (df = 3, 5, 10) or a normal mixture—dropped to 98%, 97%, 97%, and 97%, respectively. Similarly, these models experienced divergent transitions in 15%, 7%, 4%, and 7% of iterations.

4. Discussion

The results of this simulation study offer important nuance to the discussion and pursuit of adaptive shrinkage. While utilizing a Student's *t* distribution for the random effects has been theorized as a method to relax the normality assumption, the findings suggest that this added flexibility does not necessarily translate into superior classification accuracy or improved estimation of cluster-specific effects in the tails.

The investigation centered on three primary research questions regarding the classification of clusters in the tails, model robustness, and the trade-offs of adaptive shrinkage. In terms of how accurately the different models classified clusters as belonging to the tails of the distribution, using the Student's *t* approach made little difference in the accuracy of identifying clusters in the bottom or top 10% of the distribution. This is evidenced by nearly identical distributions of Brier scores across approaches and modeling conditions. In terms of robustness to violations of the normality assumption, the standard Gaussian HLM approach proved robust to violations of the normality assumption, even when the true data-generating mechanism was heavy-tailed. When the normality assumption was violated by the data-generating mechanism, the standard Gaussian HLM performed comparably to the more complex alternatives in terms of RMSE and coverage. Finally, in terms of trade-offs, while using the Student's *t* distribution did alter the mechanics of shrinkage, as seen in the differing shrinkage correlations with cluster size, it introduced some computational costs. The Bayesian Student's *t* models were more prone to divergent transitions when the degrees of freedom were not set as a fixed hyperparameter and had problems with convergence. This was particularly the case when the degrees of freedom were small.

While it is easy to view these results as disappointing, they are highly informative for both methodologists and applied researchers. Although adaptive shrinkage, as designed in this simulation, did not substantially improve classification or estimation accuracy, it did demonstrate that shrinkage is allocated differently when the variance of the true distribution of random effects is held constant.

To contextualize these findings within the broader framework of Empirical Bayes (EB) shrinkage, it is helpful to consider how the shrinkage factor behaves under different distributional assumptions. Under the standard Gaussian HLM, EB shrinkage acts as a linear estimator: the weight pulling a cluster's estimate toward the grand mean is strictly a function of its sample size (n_j) and the estimated variance components ($\hat{\sigma}^2_u$ and $\hat{\sigma}^2_e$). Consequently, shrinkage is strongly tied to cluster size, with smaller clusters uniformly experiencing greater shrinkage regardless of how extreme their specific effects, or means, might be.

By specifying a Student's *t* distribution for the random intercepts, the mechanism of EB shrinkage changes. In this study, the implementation parameterizes the Student's *t* distribution as a scale mixture of normal distributions. This means that each cluster is given a localized variance multiplier. These localized variance multipliers allow the shrinkage factors to become non-linear and adapt to the data. For instance, an extreme observed mean for a cluster can be accommodated by an inflated cluster-specific variance. This, in turn, reduces the amount of EB shrinkage applied to that outlier. The simulation confirms this shift: the Student's *t* specification substantially reduces the correlation between shrinkage and cluster size, which suggests that shrinkage was no longer strictly dictated by the cluster sample size. However, the lack of improvement in classification accuracy suggests that

while modifying the random-effects distribution successfully alters the EB shrinkage weights, this flexibility does not systematically separate true tail clusters from the rest of the distribution better than the standard Gaussian HLM.

Furthermore, these results suggest that future research on adaptive shrinkage should consider alternative methods for "turning the dial" on shrinkage. Future studies might target other components of the shrinkage equation. While the current study focused entirely on the σ_u^2 component via the choice of prior on the random effects distribution, the entire formula for the reliability factor (see equation (2)) itself and components related to the center of the distribution warrant further investigation.

This study contributes to ongoing methodological efforts to support more effective and responsible applications of hierarchical linear modeling in educational research. These efforts are particularly relevant when cluster-specific effects are estimated and used for inference. Accurate estimation and interpretation of these effects are important in contexts such as accountability systems and multisite randomized controlled trials, where cluster-specific effect estimates can inform decisions about program evaluation, implementation, and scaling. In these contexts, attention to the mechanics and implications of Empirical Bayes shrinkage is important because shrinkage affects how cluster-specific effects are estimated and subsequently interpreted for inference and decision making. The results of this study show that simply substituting a heavy-tailed parametric distribution does not resolve the potential misclassifications of extreme clusters. These findings suggest that further methodological work may benefit from exploring more flexible and context-appropriate analytical strategies that take into account the specific goals and inferential implications associated with cluster-specific effects.

5. Acknowledgments

Funding Statement The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award #R305B200026 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Competing Interests None.

6. Data and Code Availability

The Stan code used to fit the Student's *t* model and the R code used to generate the data and analyze the results of the simulation study are openly available in IMPS 2025 Proceedings Simulation Code at https://osf.io/dpsrj/overview?view_only=5aca8ed52bcc4832aacaa55d77929061, reference number 10.17605/OSF.IO/DPSRJ.

References

Arling, G., Lewis, T., Kane, R. L., Mueller, C., & Flood, S. (2007). Improving quality assessment through multilevel modeling: The case of nursing home compare. *Health Services Research*, *42*(3 Pt 1), 1177–1199. https://doi.org/10.1111/j.1475-6773.2006.00647.x

Arpino, B., & Varriale, R. (2010). Assessing the quality of institutions' rankings obtained through multilevel linear regression models. *Journal of Applied Economic Services*, *1*(11), 7–22.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. https://doi.org/10.1214/ss/1177011136

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, *3*(1), 1–10. https://doi.org/10.1080/2330443X.2016.1182878

Hui, F. K. C., Müller, S., & Welsh, A. H. (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review*, *89*(1), 186–206. https://doi.org/10.1111/insr.12378

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, *10*(2), 249–276. https://doi.org/10.1198/10618600152628059

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc.

Tong, X., & Zhang, Z. (2012). Diagnostics of robust growth curve modeling using student's t distribution. *Multivariate Behavioral Research*, *47*(4), 493–518. https://doi.org/10.1080/00273171.2012.692614

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217–221. https://doi.org/10.1080/01621459.1996.10476679

Yuan, K.-H., Bentler, P., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. https://doi.org/10.1007/BF02295644