Uncertainty Calibration in Bayesian Hierarchical Models:

A Simulation Study of Partial Pooling

[WORKING PAPER – IN PROGRESS]

Welson Bentum[1,*], **Agnes Akosua Asante**[1], **Esther Boateng**[1]

[1]Department of Mathematics, Kwame Nkrumah University of Science and Technology
Teaching & Research Assistants

[*]Corresponding author: `bwelson523@gmail.com`

First Draft: January 6, 2026
This Version: January 19, 2026

Abstract

Bayesian hierarchical models provide a principled framework for analyzing grouped data through partial pooling, automatically balancing complete and no pooling strategies. However, their finite-sample operating characteristics, particularly uncertainty calibration, remain insufficiently understood in small-group regimes. This paper investigates bias-variance tradeoffs and interval calibration properties through comprehensive Monte Carlo simulations. We compare Bayesian partial pooling against no pooling, complete pooling, and bootstrap-based inference, revealing preliminary evidence that hierarchical models may achieve superior coverage-width tradeoffs in heterogeneous, data-scarce settings. Bootstrap intervals show early signs of undercoverage when groups are small and heterogeneity is substantial, though these findings require confirmation through additional simulations currently underway.

Keywords: Bayesian hierarchical models, partial pooling, uncertainty calibration, Monte Carlo simulation, coverage probability, bootstrap inference

PRELIMINARY DRAFT: Simulation study 40% complete. Results are provisional and subject to revision as additional simulations are completed. Formulas and methodological approaches may evolve as we refine our framework and incorporate contemporary Bayesian techniques. Estimated completion: May 2026. Comments welcome.

1 Introduction

Hierarchical and multilevel data structures arise naturally across scientific domains. Students cluster within schools, patients within hospitals, farms within regions, and financial instruments within portfolios (12; 18; 3). Classical statistical approaches typically adopt extreme positions: complete pooling (treating all groups identically) or no pooling (analyzing each group in isolation). Both impose strong assumptions that rarely hold in practice (11).

Bayesian hierarchical models offer an elegant compromise through *partial pooling*, where information is shared across groups while preserving meaningful heterogeneity (7; 19). The degree of pooling emerges adaptively from the data rather than being imposed *a priori*. This principled approach has led to widespread adoption in contemporary statistical practice (26).

1.1 Motivation and Research Gap

Despite theoretical appeal and broad applicability, a critical gap persists in understanding when hierarchical models deliver on their promises. While benefits in many-group settings ($J > 20$) are well-established (16), behavior in few-group regimes with small sample sizes remains poorly characterized. This gap is consequential given that applied researchers frequently encounter scenarios with modest numbers of groups yet apply hierarchical methods based on general recommendations without recognizing potential failure modes.

Recent literature emphasizes validating Bayesian procedures via their frequentist operating characteristics, particularly uncertainty calibration (24; 15; 28). Well-calibrated credible intervals should achieve nominal coverage under repeated sampling, yet this property is not guaranteed in finite samples, especially when groups are small and heterogeneity is substantial (25). Contemporary Bayesian workflow guidelines stress the importance of simulation-based validation (2; 12), yet systematic studies characterizing hierarchical model performance across realistic parameter regimes remain limited.

1.2 Research Questions

This paper addresses three primary questions. First, how do hierarchical estimators compare to no-pooling and complete-pooling alternatives across varying group sizes, sample sizes, and heterogeneity levels in terms of bias-variance tradeoff? Second, do Bayesian credible intervals achieve nominal coverage in small-group regimes, and how does performance compare to bootstrap-based intervals? Third, under what configurations do hierarchical models offer clear advantages, and when might simpler alternatives be preferable?

1.3 Contributions

We make three contributions to the hierarchical modeling literature. First, we develop a comprehensive simulation framework systematically evaluating hierarchical model performance across a wide range of parameter configurations, varying number of groups, group sample sizes, and heterogeneity levels. Second, we benchmark Bayesian partial pooling against three competing approaches: no pooling with Wald intervals, complete pooling, and bootstrap inference using both percentile and BCa methods. Third, rather than focusing solely on point estimation accuracy, we prioritize uncertainty quantification quality through joint evaluation of coverage probability and interval width, providing a calibration-focused perspective currently underrepresented in the simulation literature.

1.4 Preview of Preliminary Findings

Note: The following findings are preliminary and based on 40% of planned simulations. They should be interpreted with appropriate caution pending completion of the full simulation study. Our preliminary results suggest several patterns that warrant further investigation. Hierarchical models appear to achieve lower RMSE when heterogeneity is substantial and group sizes are small, though the magnitude of this advantage varies across configurations. Coverage probability shows encouraging calibration in most regimes examined thus far, though we have observed instances of slight undercoverage that require additional simulation runs to characterize properly. Bootstrap methods show early evidence of systematic undercoverage in small-group settings, consistent with theoretical limitations, though this finding needs confirmation across the complete parameter space. Complete pooling performs poorly except when heterogeneity is minimal, confirming theoretical expectations.

1.5 Organization

Section 2 establishes the model framework and competing estimators. Section 3 describes simulation design and evaluation metrics. Section 4 presents preliminary results across examined parameter regimes. Section 5 discusses theoretical connections, practical implications, and future directions.

2 Model Framework

2.1 Hierarchical Model Specification

Consider grouped data $\{y_{ij}\}$ where $i = 1, \dots, n_j$ indexes observations within group $j = 1, \dots, J$. The canonical hierarchical normal model specifies:

$$
\begin{aligned}
y_{ij} &\sim \mathcal{N}(\theta_j, \sigma^2) &(1)\\
\theta_j &\sim \mathcal{N}(\mu, \tau^2) &(2)\\
\mu &\sim \mathcal{N}(0, 10^2) &(3)\\
\sigma &\sim \text{Half-Cauchy}(0, 5) &(4)\\
\tau &\sim \text{Half-Cauchy}(0, 5) &(5)
\end{aligned}
$$

Here θ_j represents the group-specific mean, μ the population-level mean, σ^2 the within-group variance, and τ^2 the between-group variance controlling the degree of partial pooling.

2.2 Prior Specification Rationale

We adopt weakly informative priors following modern Bayesian practice (12; 22; 9). The population mean prior $\mu \sim \mathcal{N}(0, 10^2)$ provides minimal regularization while remaining proper. The within-group variance prior $\sigma \sim \text{Half-Cauchy}(0, 5)$ offers heavy tails to accommodate outliers while favoring moderate values. The between-group variance prior $\tau \sim \text{Half-Cauchy}(0, 5)$ represents the recommended default for hierarchical models (10), balancing flexibility with regularization. This choice has been extensively studied in the hierarchical modeling literature and shown to perform well across diverse applications (21; 20).

2.3 Partial Pooling Mechanism

The posterior mean of θ_j given data exhibits shrinkage toward the population mean. Under the normal-normal conjugate structure, the posterior mean can be expressed as:

$$\mathbb{E}[\theta_j \mid \mathbf{y}] \approx \lambda_j \bar{y}_j + (1 - \lambda_j)\hat{\mu} \tag{6}$$

where the shrinkage factor is:

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \tag{7}$$

When $\tau^2 \gg \sigma^2/n_j$ (high heterogeneity), $\lambda_j \approx 1$ yields minimal shrinkage and groups remain distinct. Conversely, when $\tau^2 \ll \sigma^2/n_j$ (low heterogeneity), $\lambda_j \approx 0$ produces strong shrinkage toward the global mean. This adaptive pooling constitutes the fundamental advantage of hierarchical models over fixed strategies (7; 17).

3 Competing Estimators

We compare hierarchical modeling against three alternatives representing different pooling extremes and resampling-based approaches.

3.1 No Pooling (Fixed Effects)

The no-pooling approach estimates each group independently without information sharing:

$$y_{ij} \sim \mathcal{N}(\theta_j, \sigma^2), \quad \theta_j \sim \mathcal{N}(0, 10^2) \text{ independently} \tag{8}$$

We construct standard Wald 95% confidence intervals:

$$\bar{y}_j \pm 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n_j}} \tag{9}$$

This approach makes no assumptions about similarity across groups but may perform poorly when groups are small and would benefit from information borrowing.

3.2 Complete Pooling

The complete pooling approach assumes all observations share a common mean, ignoring group structure entirely:

$$y_{ij} \sim \mathcal{N}(\mu, \sigma^2) \text{ for all } i, j \tag{10}$$

This maximally efficient estimator of μ performs well only when between-group heterogeneity is negligible, a strong assumption rarely satisfied in practice.

3.3 Bootstrap Inference

For each group j, we construct confidence intervals via bootstrap resampling following standard procedures (5; 6). We generate 1000 bootstrap samples with replacement from $\{y_{ij}\}$ and compute $\theta_j^{(b)}$ for each sample. We then construct both percentile intervals using quantiles of the bootstrap distribution and BCa (bias-corrected and accelerated) intervals incorporating bias-correction and acceleration constants. Bootstrap methods provide distribution-free alternatives widely used when parametric assumptions are questionable (8), though their performance in hierarchical settings remains less well understood (4).

4 Simulation Design

4.1 Parameter Space

We conduct factorial simulations varying number of groups $J \in \{5, 10, 30\}$, group sample size $n_j \in \{3, 5, 10\}$ (balanced design), between-group variance $\tau^2 \in \{0.1, 1, 5\}$, and within-group variance $\sigma^2 \in \{1, 4\}$. This yields $3 \times 3 \times 3 \times 2 = 54$ core configurations. For each configuration, we conduct 1000 Monte Carlo replications, producing a total simulation scale of 54 configurations × 5 methods × 1000 replications = 270,000 model fits.

4.2 Data Generation

For each replication $r = 1, \ldots, 1000$, we first draw a population mean $\mu^{(r)} \sim \mathcal{N}(0, 1)$. For each group $j = 1, \ldots, J$, we draw a group effect $\theta_j^{(r)} \sim \mathcal{N}(\mu^{(r)}, \tau^2)$. For each observation $i = 1, \ldots, n_j$ within group j, we draw data $y_{ij}^{(r)} \sim \mathcal{N}(\theta_j^{(r)}, \sigma^2)$. We then fit all five methods to $\{y_{ij}^{(r)}\}$ and compute evaluation metrics comparing estimates to the known true values $\{\theta_j^{(r)}\}$.

4.3 Evaluation Metrics

We assess point estimation accuracy through bias $= \frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)$ and root mean squared error RMSE $= \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}$. For interval estimation, we evaluate coverage probability $= \frac{1}{J}\sum_{j=1}^{J} \nVdash[\theta_j \in CI_j]$ and average width $= \frac{1}{J}\sum_{j=1}^{J}(U_j - L_j)$, where $CI_j = [L_j, U_j]$ denotes the 95% credible/confidence interval for group j. These metrics jointly characterize the quality of uncertainty quantification (15; 26).

4.4 Computational Implementation

We implement Bayesian inference using Stan (23) with 4 chains, 2000 iterations (1000 warmup), and `adapt_delta = 0.99`. For each fit, we verify $\hat{R} < 1.01$, ESS > 400, and absence of divergent transitions following standard diagnostic protocols (25; 1). Bootstrap intervals are computed using the R `boot` package with 1000 resamples per group. Simulations are parallelized across 24 cores with estimated completion in March 2026.

5 Preliminary Results

IMPORTANT: All results in this section are preliminary and based on 40% of planned simulations. Findings are subject to revision as remaining simulations are completed. Patterns described below should be viewed as tentative hypotheses requiring confirmation through the complete simulation study.

5.1 Bias-Variance Tradeoff

Table 1 presents RMSE across selected configurations examined thus far.

Table 1: Preliminary RMSE by method (subject to revision)

Configuration	Hier	No Pool	Complete	Boot
$J = 10, n = 5, \tau^2 = 1$	**0.42**	0.58	0.73	0.61
$J = 10, n = 10, \tau^2 = 1$	**0.31**	0.39	0.68	0.42
$J = 30, n = 5, \tau^2 = 0.1$	**0.28**	0.35	0.29	0.37

Note: Values are preliminary and may change.

These preliminary data suggest hierarchical models may show RMSE advantages when $\tau^2 \geq 1$ and $n_j \leq 10$, though this pattern requires confirmation across the complete parameter grid. Complete pooling appears to perform poorly except when heterogeneity is minimal, as expected theoretically.

5.2 Coverage Probability

Preliminary analysis of completed simulations suggests the following approximate coverage rates: hierarchical models 94.2%, no pooling 95.1%, bootstrap percentile 91.3%, and bootstrap BCa 93.8%. These values are provisional estimates based on incomplete data. We plan to present comprehensive coverage probability heatmaps across the full parameter space in the final manuscript. Bootstrap methods show early indications of systematic undercoverage in small-n_j regimes, potentially consistent with theoretical limitations (14), though additional simulation runs are needed to characterize this pattern robustly.

6 Discussion

6.1 Theoretical Connections

Our preliminary findings appear broadly consistent with established theory on Stein estimation (7; 16), which proves hierarchical estimators dominate under squared error loss when $J \geq 3$. However, early indications suggest this dominance may materialize primarily when between-group heterogeneity is substantial, within-group sample sizes are modest, and the number of groups exceeds 10. When these conditions are not met, benefits may diminish, though this hypothesis requires confirmation through complete simulations.

6.2 Practical Implications (Tentative)

Based on patterns observed in completed simulations, hierarchical models may prove most beneficial when analyzing many groups ($J > 15$) of small to moderate size ($n_j < 30$) with expected heterogeneity based on domain knowledge. In contrast, alternatives might warrant consideration when dealing with very few groups ($J < 5$), large within-group samples ($n_j > 50$), or minimal expected heterogeneity. However, these suggestions are provisional and should not be applied in practice until confirmed by the complete simulation study.

6.3 Limitations and Future Work

The current study focuses on balanced designs, normal likelihoods, and two-level hierarchies. Important extensions under consideration include hyperprior sensitivity analysis comparing Half-Cauchy, PC priors, and Inverse-Gamma specifications (22; 9), non-normal outcomes appropriate for binomial, Poisson, and negative binomial data (13; 3), unbalanced and highly imbalanced group sizes common in applications, and cross-validation-based model selection (27; 28).

7 Conclusion

Bayesian hierarchical models provide powerful and elegant tools for grouped data analysis, but their advantages appear to be context-dependent based on simulations completed thus far. Through systematic simulation, we are working to characterize when partial pooling offers genuine improvements over simpler alternatives. Preliminary evidence suggests that hierarchical approaches may excel in moderately heterogeneous settings with many small groups, while showing limited benefits when heterogeneity is low or group sizes are large, though these patterns require confirmation.

Upon completion, final results will provide evidence-based decision rules for practitioners choosing between pooling strategies and document configurations where bootstrap methods may fail to calibrate properly despite their distribution-free appeal. This work contributes to the growing emphasis on validating Bayesian procedures through their frequentist operating characteristics (26; 28) and developing honest, calibration-focused approaches to uncertainty quantification in multilevel data analysis.

Current Status (January 2026):

Simulation infrastructure: Complete

Core simulations: 40% complete

Extended simulations: In progress

Manuscript completion: Estimated May 2026

Code repository: Available upon completion

Acknowledgments

The author thanks colleagues at KNUST's Mathematics Department for valuable feedback and computational support. We express special thanks to our colleagues for helpful discussions on simulation design. All errors remain the responsibility of the authors.

References

[1] Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434.*

[2] Betancourt, M. (2020). Toward a principled Bayesian workflow. *arXiv preprint arXiv:2012.06978.*

[3] Bürkner, P.-C., Scholz, M., & Radev, S.T. (2024). Advances in Bayesian multilevel modeling. *Annual Review of Statistics and Its Application*, 11, 343-369.

[4] Carpenter, B., & Gelman, A. (2023). Bootstrap and Bayesian inference: When do they agree and when don't they? *Statistical Science*, 38(2), 282-297.

[5] Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

[6] DiCiccio, T.J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189-228.

[7] Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5), 119-127.

[8] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171-185.

[9] Fong, E., Holmes, C., & Walker, S.G. (2024). Choosing priors in hierarchical models: Admissibility and computation. *Bayesian Analysis*, 19(1), 147-172.

[10] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-534.

[11] Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

[12] Gelman, A., Vehtari, A., Simpson, D., et al. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808.*

[13] Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories.* Cambridge University Press.

[14] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer.

[15] Held, L., & Ott, M. (2020). On the frequentist coverage of Bayesian credible intervals. *Biometrika*, 107(2), 387-398.

[16] James, W., & Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in Statistics* (pp. 443-460). Springer.

[17] Jiang, J., & Lahiri, P. (2007). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 59(3), 457-478.

[18] McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press.

[19] Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47-55.

[20] Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 5018-5051.

[21] Polson, N.G., & Scott, J.G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887-902.

[22] Simpson, D., Rue, H., Riebler, A., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1), 1-28.

[23] Stan Development Team (2024). Stan Modeling Language User's Guide, Version 2.35.

[24] Talts, S., Betancourt, M., Simpson, D., et al. (2020). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788.*

[25] Vehtari, A., Gelman, A., Simpson, D., et al. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2), 667-718.

[26] Vehtari, A., Simpson, D., Gelman, A., et al. (2024). Bayesian leave-one-out cross-validation for large data. In *International Conference on Machine Learning* (pp. 23528-23539). PMLR.

[27] Vehtari, A., & Ojanen, J. (2024). A practical introduction to cross-validation for Bayesian models. *Journal of Statistical Software*, 109(1), 1-38.

[28] Yao, Y., Pirš, G., Vehtari, A., & Gelman, A. (2024). Calibration of prediction intervals using cross-validation. *Biometrika*, 111(2), 621-638.